PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12Q 1/68, C12P 19/34 C07H 15/12, G01N 33/53

A1

(11) International Publication Number: WO 92/13969

(43) International Publication Date: 20 August 1992 (20.08.92)

(21) International Application Number: PCT/US92/00736

(22) International Filing Date: 28 January 1992 (28.01.92)

(30) Priority data: 647,655 31 January 1991 (31.01.91) US

(71) Applicant: BAYLOR COLLEGE OF MEDICINE [US/US]; One Baylor Plaza, Houston, TX 77030 (US).

(72) Inventors: CASKEY, C., Thomas ; 6402 Belmont, Houston, TX 77005 (US). EDWARDS, Albert, O. ; 2322 Camden #2, Houston, TX 77021 (US).

(74) Agent: PAUL, Thomas, D.; Fulbright & Jaworski, 1301 McKinney, Suite 5100, Houston, TX 77010-3095 (US).

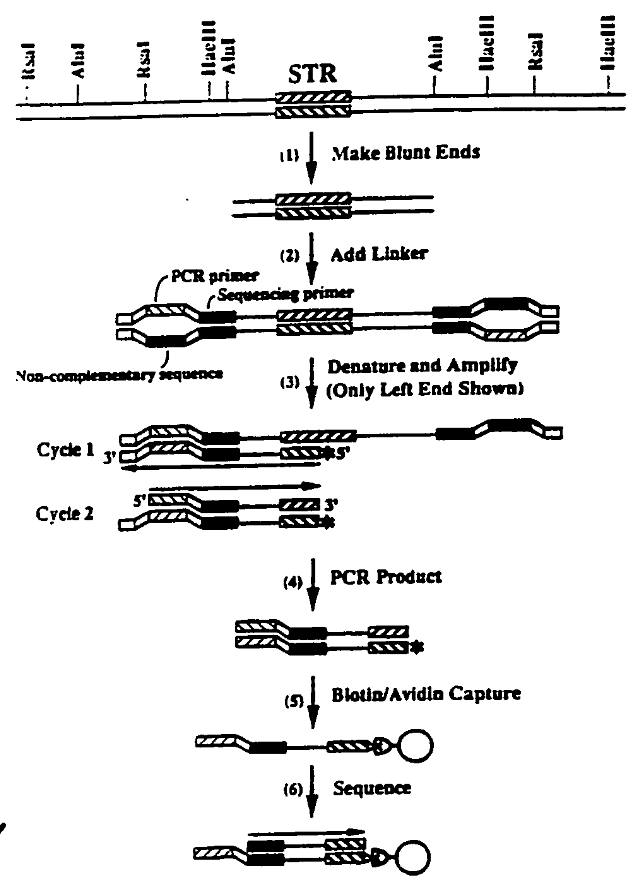
(81) Designated States: AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent).

Published
With international search report.

(54) Title: DNA TYPING WITH SHORT TANDEM REPEAT POLYMORPHISMS AND IDENTIFICATION OF POLYMORPHIC SHORT TANDEM REPEATS

(57) Abstract

The present invention relates to a DNA assay for detecting polymorphisms. Method includes the steps of extracting DNA from a sample to be tested, amplifying the extracted DNA and identifying the amplified extension products for each different sequence. Each different sequence is differentially labeled. A short tandem repeat sequence which can be characterized by the formula $(A_wG_xT_yC_z)_n$, wherein A, G, T and C represent the nucleotides, w, x, y and z represent the number of nucleotide and range from 0 to 7 and the sum of w+x+y+z ranges from 3 to 7 and n represents the repeat number and ranges from 5 to 50.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FI	Finland	ML	Mali
AU	Australia	FR	France	MN	Mongolia
BB	Barbados	GA	Gabon	MR	Mauritania
BE	Belgium	GB	United Kingdom	MW	Malawi
BF	Burkina Faso	GN	Guinea	NL	Netherlands
BG	Bulgaria	GR	Greece	NO	Norway
BJ	Benin	HU	Hungary	PL	Poland
BR	Brazil	IE	Ireland	RO	Romania
CA	Canada	IT	Italy	RU	Russian Federation
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SN	Senegal
CI	Côte d'Ivoire	LI	Liechtenstein	SU	Soviet Union
CM	Cameroon	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
DE	Germany	MC	Monaco	US	United States of America
DK	Denmark	MG	Madagascar		
ES	Spain				

DNA TYPING WITH SHORT TANDEM REPEAT POLYMORPHISMS AND IDENTIFICATION OF POLYMORPHIC SHORT TANDEM REPEATS

Field of the Invention

The present invention relates generally to a method of DNA typing for the detection of short tandem repeat sequence polymorphisms. More particularly, it relates to the method of detecting short tandem repeat sequences which show polymorphisms in the number of repeats for the detection or identification of medical and forensic samples, paternity, sample origin and tissue origin. Additionally it further relates to the method of identifying polymorphic short tandem repeats in genomes.

Background of the Invention

The volume of crime committed in the United States has risen with the increase of population and expansion of population centers. A large portion of violent crimes involve the creation of body fluid evidence having the potential of providing significant information about the perpetrator of a particular offense. Although the forensic science community has made tremendous effort in using this evidence, the results have historically been limited and are not useful in many situations when dealing with human

remains and crime scene evidence. Identification by genetically inherited markers has long been seen as a possibility that would overcome most of the problems that are encountered when identification is not accomplished by fingerprints, forensic odontology, medical records or other methods. The establishment of a genetically inherited method that could be used for identification would have tremendous impact on investigation of the violent crimes of sexual assault and murder, identification of human remains and missing persons, and disputed parentage.

Methods enabling the matching of unidentified tissue samples to specific individuals would have wide application in the criminal justice system and the forensic sciences. With the possible exception of monozygotic twins, each individual in the human population has a unique genetic composition which could be used to specifically identify each individual. This phenomenon presents the theoretical possibility of using DNA sequence variation to determine whether a forensic sample was derived from any given individual.

Genetic marker systems, including blood groups and isoenzymes, have been used by forensic and medical serologists to provide estimates of individuality ranging from 1:1000 to 1:1,000,000 using 10 to 15 markers. Numbers in this range are often not available since a large percentage of the evidence does not yield results for ten genetic marker systems. Forensic scientists, investigators and the court system have been using inclusions as low as 1:5 to 1:100 in a population to bolster their case against defendants.

The fields of forensic and medical serology, paternity testing, and tissue and sample origin has been altered by the use of DNA sequence variation, e.g., satellite sequences and variable number of tandem repeats (VNTRS) or AMP-FLPS, in the crime laboratory, the court, hospitals and research and testing labs. Inclusion probabilities stated by the laboratories performing the analyses in such cases often exceed 1:1,000,000. The first implementation of DNA typing in forensics, was Jeffreys' use of a multilocus DNA probe "fingerprint' that identified a suspect in a murder case occurring in England. In the United States, DNA profiling has been established using a battery of unlinked highly polymorphic single locus VNTR probes. The use of these batteries of probes permits the development of a composite DNA profile for an individual. These profiles can be compared to ethnic databases using the principles of Hardy-Weinberg to determine the probability of the match between suspect and unknown forensic samples.

The application of VNTRs to gene mapping, population genetics, and personal identification has been limited by the low frequency and asymmetric distribution of these repeats in the genome and by the inability to precisely determine alleles with Southern hybridization-based detection schemes. The inability to make precise allele determinations complicates the application of VNTRs to personal identification. Binning protocols have been devised in which all alleles occurring within a region of the gel are treated as the same allele for genotype calculations. Since the allele distribution appears continuous because of the limited resolving power of Southern

gels, heterozygotes with alleles of similar size may be scored as homozygotes. These features have led to claims that VNTR loci are not in Hardy-Weinberg equilibrium, and therefore the method for calculating the significance of a match is not agreed upon.

Although these methods have markedly improved the power of the forensic and medical scientist to distinguish between individuals, they suffer from a number of shortcomings including a lack of sensitivity, the absence of internal controls, expense, time intensity, relatively large sample size, an inability to perform precise allele identification and problems with identifying degraded DNA samples.

Medical and forensic studies have also employed the polymerase chain reaction (PCR) to examine variation in the HLA locus. PCR has also been used to amplify short VNTRs or AMP-FLPs. The use of PCR addresses some of the problems of sensitivity and sample degradation, however, the HLA typing system, still has some problems. A simpler, more powerful technique is needed which makes use of the most recent advances in DNA technology.

The present invention involves the novel application of these advances to medical and forensic science. In the present invention novel classes of highly polymorphic, primarily trimeric and tetrameric, short or simple tandem repeats (STRs) which are present within the human genome have been identified. These STRs have characteristics suitable for inclusion in a DNA profiling assay. This assay incorporates internal or external standards, provides higher sensitivity,

requires shorter analysis time, lowers expense, and enables precise identification of alleles. The STRs are amplified with great fidelity and the allele patterns are easily interpreted. Amplification of highly polymorphic tandemly reiterated sequences may be the most cost effective and powerful method available to the medical and forensic community.

The DNA profiling assay of the present invention has features which represent significant improvements over existing technology and brings increased power and precision to DNA profiling for criminal justice, paternity testing, and other forensic and medical uses.

Summary of the Invention

An object of the present invention is a method for DNA profiling using short tandem repeat polymorphisms.

An additional object of the present invention is a method for identifying the source of DNA in a forensic or medical sample.

A further object of the present invention is to provide an automated DNA profiling assay.

An additional object of the present invention is the provision of a method for identifying and detecting short tandem repeat polymorphisms to expand the discriminating power of a DNA profiling assay.

A further object of the present invention is to extend the discriminating power of a DNA profiling assay.

An additional object of the present invention is to provide a kit for detecting short tandem repeat polymorphisms.

Thus in accomplishing the foregoing objects, there is provided in accordance with one aspect of the present invention a DNA profiling assay comprising the steps of: extracting DNA from a sample to be tested; performing multiplex polymerase chain reaction on the extracted DNA; and identifying the amplified extension products from the multiplex polymerase chain reaction for each different sequence, wherein each different sequence is differentially labelled.

The DNA profiling assay is applicable to any sample from which amplifiable DNA can be extracted. In medical and forsenic uses the samples are selected from the group consisting of blood, semen, vaginal swabs, tissue, hair, saliva, urine and mixtures of body fluids.

Specific embodiments of the invention include the use of short tandem repeat sequences selected from the group of non-duplicative nucleotide sequences consisting of:

$(AA)_m$, $(AC)_m$, $(AG)_m$, $(AT)_m$, $(CC)_m$, $(CG)_m$,
$(AAC)_n$, $(AAG)_n$, $(AAT)_n$, $(ACC)_n$, $(ACG)_n$,
$(ACT)_n$, $(AGC)_n$, $(AGG)_n$, $(ATC)_n$, $(CCG)_n$,
$(AAAC)_n$, $(AAAG)_n$, $(AAAT)_n$, $(AACC)_n$, $(AACG)_n$, $(AACT)_n$,
$(AAGC)_n$, $(AAGG)_n$, $(AAGT)_n$, $(AATC)_n$, $(AATG)_n$, $(AATT)_n$,
$(ACAG)_n$, $(ACAT)_n$, $(AGAT)_n$, $(ACCC)_n$, $(ACCG)_n$, $(ACCT)_n$,
$(ACGC)_n$, $(ACGG)_n$, $(ACGT)_n$, $(ACTC)_n$, $(ACTG)_n$, $(ACTT)_n$,

$(AGCC)_n$, $(AGCG)_n$, $(AGCT)_n$, $(AGGC)_n$, $(AGGG)_n$, $(ATCC)_n$, $(ATCG)_n$, $(ATGC)_n$, $(CCCG)_n$, $(CCGG)_n$ and combinations thereof wherein n and m are the repeat number and m varies from about 10 to 40 and n varies from about 5 to 40.

In another embodiment of the present invention the differential label for each specific sequence is selected from the group consisting of fluorescers, radioisotopes, chemiluminescers, enzymes, stains and antibodies. One specific embodiment uses the fluorescent compounds Texas Red, tetramethylrhodamine-5-(and-6) isothiocyanate, NBD aminoheanoic acid and fluorescein-5-isothiocyanate.

The assay can be automated by using an automated fluorescent DNA label analyzer capable of distinguishing, simultaneously, different fluors during the identifying step.

Another embodiment of the present invention includes a kit containing oligonucleotide primers for the short tandem repeat sequences.

A further embodiment includes a method for detecting polymorphic short tandem repeats comprising the steps of: determining non-duplicative nucleotide sequences of the formula $(A_wG_xT_yC_z)$ wherein A,G,T, and C represent the nucleotides; and w, x, y and z represent the number of each nucleotide in the sequence and range between 0 and 7 with the sum of w+x+y+z ranging from 3 to 7; identifying and searching for $(A_wG_xT_yC_z)_n$ in databases containing known genetic sequences, wherein n represents the number of tandem repeats of the sequence and is at least about 5; extracting each nucleotide sequence and its flanking sequences found in the searching step; identifying the extracted sequences

which have unique flanking sequences; synthesizing oligonucleotide primer pairs corresponding to the flanking sequences; performing PCR with the primer pairs on DNA samples from a test population; and examining the extension products from the PCR to detect polymorphic short tandem repeats.

Other and further objects, features and advantages will be apparent and the invention will be more readily understood from a reading of the following specification and by reference to the accompanying drawings, forming a part thereof, where examples of the presently preferred embodiments of the invention are given for the purpose of disclosure.

Description of the Drawings

Figure 1 illustrates the strategy used to determine the sequence flanking a STR.

Figure 2 shows the development of a polymorphic STR locus.

Figure 3 shows examples of products from the multiplex and single PCR assays used in generating multilocus genotype data. Figure 3A shows the multiplex PCR(mPCR) of HUMHPRTB$[AGAT]_n$ (top) and HUMFABP$[AAT]_n$. Figure 3B shows the mPCR of HUMRENA4$[ACAG]_n$ (top) and HUMTH01$[AATG]_n$. Figure 3C shows the single PCR of HUMARA$[AGC]_n$.

Figures 4A to 4E plot the relative allele frequency distributions. Allele counts were used to calculate and plot the relative frequencies of alleles at: 4A HUMRENA4$[ACAG]_n$; 4B HUMTH01$[AATG]_n$; 4C HUMARA$[AGC]_n$; 4D HUMHPRTB$[AGAT]_a$; and 4E HUMFABP$[AAT]_n$.

Figure 5 shows the results of a fluorescent DNA typing assay. The analysis software scales the intensities of the fluorescent profiles relative to the strongest signal.

The drawings and figures are not necessarily to scale and certain features of the invention may be exaggerated in scale or shown in schematic form in the interest of clarity and conciseness.

Detailed Description of Invention

It will be readily apparent to one skilled in the art that various substitutions and modifications may be made to the invention disclosed herein without departing from the scope and the spirit of the invention.

As used herein, the term "short tandem repeat" (STR) refers to all sequences between 2 and 7 nucleotides long which are tandemly reiterated within the human organism. The STRs can be represented by the formula $(A_wG_xT_yC_z)_n$ where A,G,T an C represent the nucleotides which can be in any order; w, x, y and z represent the number of each nucleotide in the sequence and range between 0 and 7 with the sum of w+x+y+z ranging between 2 and 7; and n represents the number of times the sequence is tandemly repeated and is between about 5 and 50. Most of the useful polymorphisms usually occur when the sum of w+x+y+z ranges between 3 and 7 and n ranges between 5 and 40. For dimeric repeat sequences n usually ranges between 10 and 40.

As used herein "non-duplicative" sequence means the sequence and its complement. It is represented in its lowest alphabetical form as shown in Table 1. For example (ACT) represents ACT, CTA, TAC, AGT, TAG and

GTA. Each representative sequence can represent a maximum of two times the number of nucleotides in the sequence.

As used herein "flanking sequence" refers to the nucleotide sequence on either side of the STR. "Unique flanking sequences" are those flanking sequences which are only found at one location within the genome.

The term "oligonucleotide primers" as used herein defines a molecule comprised of more than three deoxyribonucleotides or ribonucleotides. Its exact length will depend on many factors relating to the ultimate function and use of the oligonucleotide primer, including temperature, source of the primer and use of the method. The oligonucleotide primer can occur naturally, as in a purified restriction digest, or be produced synthetically. The oligonucleotide primer is capable of acting as an initiation point for synthesis when placed under conditions which induce synthesis of a primer extension product complementary to a nucleic acid strand. The conditions can include the presence of nucleotides and an inducing agent such as a DNA polymerase at a suitable temperature and pH. In the preferred embodiment, the primer is a single-stranded oligodeoxyribonucleotide of sufficient length to prime the synthesis of an extension product from a specific sequence in the presence of an inducing agent. Sensitivity and specificity of the oligonucleotide primers are determined by the primer length and uniqueness of sequence within a given sample of template DNA. In the present invention the oligonucleotide primers are usually about greater than 15 mer and in the preferred embodiment are about 20 to 30 mer in length.

Each pair of primers is selected to detect a different STR. Each primer of each pair herein is selected to be substantially complementary to a different strand in the flanking sequence of each specific STR sequence to be amplified. Thus one primer of each pair is sufficiently complementary to hybridize with a part of the sequence in the sense strand and the other primer of each pair is sufficiently complementary to hybridize with a different part of the same sequence in the antisense strand. Although the primer sequence need not reflect the exact sequence of the template, the more closely the 3′ end reflects the exact sequence, the better the binding during the annealing stage.

As used herein the term "extension product" refers to the nucleotide sequence which is synthesized from the 3′ end of the oligonucleotide primer and which is complementary to the strand to which the oligonucleotide primer is bound.

As used herein the term "differentially labeled" indicates that each extension product can be distinguished from all others because it has a different label attached and/or is of a different size and/or binds to a specifically labeled oligonucleotide. One skilled in the art will recognize that a variety of labels are available. For example, these can include radioisotopes, fluorescers, chemiluminescers, stains, enzymes and antibodies. Various factors affect the choice of the label. These include the effect of the label on the rate of hybridization and binding of the primer to the DNA, the sensitivity of the label, the ease of making the labeled primer, probe or extension products, the ability to automate, the available instrumentation, convenience and the like. For example, differential radioisotope labelling could

include ^{32}P, ^{3}H and ^{14}C; differential fluorescers labelling could include fluorescein-5-isothiocyanate, tetramethylrhodamine-5- (and-6) isothiocyanate, Texas Red and NBD aminoheanoic acid; or a mixture of different labels such as radioisotopes, fluorescers and chemiluminescers.

Each specific, different DNA sequence, which is to be detected herein is derived from genomic DNA. The source of the genomic DNA to be tested can be any medical or forensic sample. Examples of medical and forensic samples include blood, semen, vaginal swabs, tissue, hair, saliva, urine and mixtures of body fluids. These samples can be fresh, old, dried and/or partially-degraded. The samples can be collected from evidence at the scene of a crime.

The term "forensic sample" as used herein means using the technology for legal problems including but not limited to criminal, paternity testing and mixed-up samples. The term "medical sample" as used herein means use of the technology for medical problems including but not limited to research, diagnosis, and tissue and organ transplants.

As used herein the term "polymorphism" refers to the genetic variation seen in the tandem repeats or flanking sequences. One example of this polymorphism is in the number of times the 3 to 7 nucleotide sequence is repeated.

As used herein the term "multiplex polymerase chain reaction" (mPCR) refers to a novel variation of PCR. It is a procedure for simultaneously performing PCR on greater than two different sequences. The mPCR reaction comprises: treating said extracted DNA to form single stranded complementary strands, adding a plurality of labeled paired oligonucleotide primers,

each paired primer specific for a different short tandem repeat sequence, one primer of each pair substantially complementary to a part of the sequence in the sense strand and the other primer of each pair substantially complementary to a different part of the same sequence in the complementary antisense strand, annealing the plurality of paired primers to their complementary sequences, simultaneously extending said plurality of annealed primers from the 3′ terminus of each primer to synthesize an extension product complementary to the strands annealed to each primer, said extension products, after separation from their complement, serving as templates for the synthesis of an extension product for the other primer of each pair, separating said extension products from said templates to produce single stranded molecules, amplifying said single stranded molecules by repeating at least once said annealing, extending and separating steps. In the mPCR process the preferred method for three loci includes: (1) primers composed of similar GC base compositions and lengths; (2) longer extension times up to 8 fold the normally utilized times; and (3) minimization of the number of PCR cycles performed to achieve detection for example approximately 23-25 cycles.

The mPCR reaction is optimized for each reaction. In some mPCR reactions the optimization further includes more enzyme than one normally adds to a PCR reaction.

As used herein, the term "match probability" refers to the chance that two unrelated persons will have the same combined genotype at the examined loci.

One embodiment of the present invention is a DNA profiling assay for detecting polymorphisms in a short

tandem repeat, comprising the steps of: extracting DNA from a sample to be tested; amplifying the extracted DNA; and identifying said amplified extension products for each different sequence, wherein each different sequence has a differential label. Although a variety of known amplification procedures are known, the preferred embodiment employs PCR or mPCR.

In one embodiment of the present invention an external standard is used. In a preferred embodiment an internal standard is used. The internal standard is composed of labeled alleles of the STR loci of interest. One skilled in the art will recognize that the choice of standard and the intervals chosen will depend on the label and the desired resolution. For example, in a DNA profiling assay STR alleles from a DNA sample can be localized with greater than 1 bp resolution using an internal standard marker every 3-5 bp.

In a preferred assay, short tandem repeat sequences of nucleotides characterized by the formula $(A_wG_xT_yC_z)_n$ wherein A, G, T and C represent the nucleotides, w, x, y ans z represent the number of each nucleotide and range from 0 to 7 and the sum of w+x+y+z ranges from 3 to 7 and n represents the repeat number and ranges from 5 to 40 are used. In another preferred assay, the sum of w+x+y+z is either 3 or 4. In the preferred embodiment of the profiling assay, at least two STRs are assayed simultaneously.

In the most preferred embodiment the DNA profiling assay is automated. This automation can be achieved by a variety of methods. One method is to use an automated DNA label analyzer capable of distinguishing simultaneously different fluorescers, radioactive labels or chemiluminescers during the identifying step.

One skilled in the art will readily recognize that a variety of instrumentation meets this requirement. One example of such an analyzer used in the preferred assay is the Applied Biosystems 370A Fluorescent DNA sequence device ("370A device") which has the capability to distinguish between four different fluors during electrophoresis.

Another aspect of the present invention is a method of detecting polymorphic STRs for use in the DNA profiling assay. The method of detecting a polymorphic STR comprises the steps of: determining possible non-duplicative nucleotide sequences of the formula $(A_wG_xT_yC_z)$, wherein A, G, T and C represent each respective nucleotide and w, x, y and z represent the number of each nucleotide in the sequence and range between 0 and 7 with the sum of w+x+y+z ranging between 2 and 7; searching for and identifying $(A_wG_xT_yC_z)_n$ in databases containing known genetic sequences and identifying the $(A_wG_xT_yC_z)_n$ sequence of said genetic sequence and its flanking sequence, wherein n represents the number of tandem repeats of the sequence and is at least about 5; extracting each identified sequence and its flanking sequence; identifying the extracted sequences which have unique flanking sequences; synthesizing oligonucleotide primer pairs to the unique flanking sequences; performing a PCR with the primer pairs on DNA samples from a test population; and examining the extension products of the PCR to detect polymorphic STRs.

A further aspect of the present invention is the provision of a kit for DNA profiling assays. The kit is comprised of a container having an oligonucleotide primer pairs for amplifying a STR. In the preferred embodiment the number of STR primer pairs is selected

such that the genotype frequency (p) is at least 10^{-6}. This usually requires 6-10 STR primer pairs.

A further addition to the kit can be a container having labelled standards. An additional enhancement to the kit is the addition of reagents for mPCR.

The following examples are offered by way of illustration and are not intended to limit the invention in any manner. In examples, all percentages are by weight, if for solids and by volume if for liquids, and all temperatures are in degrees Celsius unless otherwise noted.

Example 1

Computer Identification of STR loci

STRs were identified by searching all human sequences in the GenBank DNA sequence repository for the presence of all possible classes of dimeric, trimeric, and tetrameric STRs. One skilled in the art will readily recognize that a similar search using repeats of 5 to 7 nucleotides can be used to identify STRs of 5 to 7 nucleotides in length. The possible non-duplicative nucleotide sequences used in this search are given by their lowest alphabetical representation.

Table 1: Possible Non-Duplicative STRs

(AA)	(AC)	(AG)	(AT)	(CC)	(CG)		
(AAC)	(AAG)	(AAT)	(ACC)	(ACG)	(ACT)	(AGC)	(AGG)
(ATC)	(CCG)	(AAAC)	(AAAG)	(AAAT)	(AACC)	(AACG)	(AACT)
(AAGC)	(AAGG)	(AAGT)	(AATC)	(AATG)	(AATT)	(ACAG)	(ACAT)
(AGAT)	(ACCC)	(ACCG)	(ACCT)	(ACGC)	(ACGG)	(ACGT)	(ACTC)

(ACTG) (ACTT) (AGCC) (AGCG) (AGCT) (AGGC) (AGGG) (ATCC) (ATCG) (ATGC) (CCCG) (CCGG)

As shown in Table 2, the computer search identified a considerable number of STRs. The dimeric search was set to identify sequences in which the STR was repeated at least 10 times and the trimeric and tetrameric search was set to identify sequences which were repeated at least 5 times.

TABLE 2: HUMAN STRs IN THE GENBANK DNA

MONOMER LENGTH	DINUCLEOTIDE (10 OR MORE)	TRINUCLEOTIDE (5 OR MORE)	TETRANUCLEOTIDE (5 OR MORE)
FRACTION OBSERVED	5/6	10/10	16/34
TOTAL NUMBER	217	101	67

The fraction observed refers to the number of different classes of STRs observed out of the total number of possible STR classes. All classes of trimeric repeats were present and about half of the possible tetrameric sequences were represented.

Approximately 50% of STRs studied were polymorphic. Trimeric and tetrameric STRs have features of polymorphic markers useful for the physical and genetic mapping of the human genome and personal identification in the medical and forensic sciences.

Example 2

Molecular Biological Identification of STR Loci

In addition to the procedure for identifying STRs in the GenBank, other methods are available to identify additional STR loci. For example, oligonucleotide probes for the possible 50 unique dimeric, trimeric and

tetrameric STR sequences can be synthesized and used to screen total human DNA libraries. In one example recombinant bacteriophage lambda of the human X chromosome were plated at a density of 255 plaque forming units per 15 cm plate. Plaque lifts made from the plates are hybridized to ^{32}P 5′ end-labeled oligonucleotides of the STR motifs. Standard hybridization methods were used. Oligonucleotides were labeled according to standard protocols.

With nucleotide sequences up to above 100 bp the conditions for hybridization may be estimated using the following formula:

$$T_i = T_m - 15°C$$

$$T_m = 16.6log[M] + 0.41[P_{gc}] + 81.5 - P_m - B/L - 0.65[P_f]$$

where: M is the molar concentration of Na^+, to a maximum of 0.5 (1 X SSC contains 0.165 M Na^+);

P_{gc} is the percent of G or C bases in the oligonucleotide and is 1-16between 30 and 70;

P_m is the percent of mismatched bases, if known;

P_f is the percent of formamide in the buffer;

B is 675 for synthetic probes up to 100 bases;

L is the length of the probe in bases.

The formula was used to arrive at the conditions in Table 3:

Table 3

Oligo Id	sequence	(ml)2X Hyb Mix	P_f(%)	Formamide(ml)	H_2O(ml)	V(ml)
1152	$[AATC]_{7.5}$	12.5	10%	2.5	10	25
1154	$[AGAT]_{7.5}$	12.5	10	2.5	10	25
1525	$[AAT]_{10}$	12.5	0	0	12.5	25
1526	$[AATG]_{7.5}$	12.5	10	2.5	10	25
1528	$[ACAG]_{7.5}$	12.5	30	7.5	5	25

The 2 X hybridization mix contained 37.5 ml of 20 X SSC, 15 ml of 50 X Denhardts, 7.5 ml of 20% SDS, and 15 ml of H_2O. Hybridizations were performed at 42°C.

These conditions were used to determine the frequency of each STR shown on the X chromosome using recombinant lambda from an X chromosome genomic library picked to a grid. (Table 4).

TABLE 4. The frequency of trimeric and tetrameric STRs.

STR	Positive bacteriophage(%)	Frequency (kb/STR)
[AAT]	5	300
[AATC]	5	300
[AATG]	3	500
[ACAG]	3	500
[AGAT]	3	500

A total of 1020 recombinant bacteriophage were hybridized to radiolabeled 30 bp oligonucleotides (e.g., $[AATC]_{7.5}$). Calculations were based on an average insert size of 15 kb in the library. These hybridization results and the results from the GenBank studies suggest the presence of approximately 400 million STRs in the human genome. The X-chromosome results have been extended to the entire human genome by utilizing the complete genomic phage lambda library. Thus identification of sufficient STRs to extend the

DNA typing assay to very high levels of individualization (e.g., one in a billion) is feasible.

Example 3

Determination of DNA sequence flanking STRs

Clones containing STRs, for example M13, lambda, cosmid and YAC, can be identified by any procedure which allows hybridization to one of the core oligonucleotides. Most of the hybridization methods are usually laborious for determining the sequence of the unique DNA segments flanking both sides of the STR. In the present invention a strategy called STR-PCR was used. This strategy is shown in schematic form in (Figure 1).

The STR-PCR strategy was based upon the method of Riley, et al, Nucleic Acids Res. 18: 2887 (1990). The Riley method was designed to amplify the ends of YAC molecules from total yeast genomic DNA. This method was adapted to amplify the DNA segments flanking STRs, and coupled to direct DNA sequencing of the products via a solid-phased-DNA sequencing technology. The procedure involves the following steps: (1) Blunt ends are generated to flank both sides of a STR in a cloned DNA segment by digestion with a single restriction enzyme. Multiple enzymes can be used, separately, to generate a flanking sequence length in the range of 100 to 150 bp. (2) A linker which contains a region of non-complementary DNA is ligated to the population of blunt ended molecules. (3) The flanking sequences are amplified in separate reactions. The left end is amplified with the anchored PCR primer and a primer of one strand of the STR. The right end is amplified with the same anchored PCR primer and a primer of the other strand of the STR. The STR primers may be biotinylated

(*). (4) The final biotinylated (*) PCR product. (5) The biotinylated strand may be captured with avidin coated beads. And (6) The flanking sequence may be obtained by extension from the sequencing primer in the presence of dideoxynucleotides.

Results from using the STR-PCR strategy are shown in Figure 2. Amplification of the DNA sequence flanking both sides of an (AGAT) STR from two recombinant bacteriophage is shown in Figure 2A. While Figure 2B shows direct DNA sequencing of single stranded template following capture and strand separation of the biotinylated amplification products of λAE[AGAT]-2 with avidin coated magnetic beads. Figure 2C demonstrates the use of oligonucleotides complementary to the sequence flanking the STR to amplify the STR locus in a family.

Oligonucleotide primers which generate a blunt ended linker upon annealing were synthesized. Examples of these oligonucleotides are SEQ ID Nos: 1, 2, 3 and 4. Oligonucleotides (SEQ ID Nos: 1 and 2) form the double stranded linker, oligonucleotide (SEQ ID No: 3) is the PCR primer for the anchor and oligonucleotide (SEQ ID No: 4) is the DNA sequencing primer.

Oligonucleotide (SEQ ID No: 1) was phosphorylated and annealed to (SEQ ID No: 2) to form the double stranded linker by standard protocol. In this procedure 10μL of one of the linker oligonucleotides (SEQ. ID No: 1) (100μM; 1 nanomole) is combined with about 10μL of ^{32}P-gamma-ATP (10mM); about 5μL of PNK buffer; about 3μL of T4 PNKinase (30U); about 22μL of H_2O for about a 50μL final volume (at about 37°C, for about 40 min.; and at about 65°C, for about 5 min). Then about 10μL of the other linker oligonucleotide (SEQ ID No: 2) (100μM) is added. This mixture is held

at about 95°C for about 5 min. The mixture is then slowly cooled to room temperature.

Although the Riley method taught that both oligonucleotides should be phosphorylated, the present invention has discovered that it is sufficient and possibly better to phosphorylate only the first oligonucleotide.

The next procedure was to ligate the double stranded linker (PCR anchor) to recombinant bacteriophage lambda DNA. First each clone was digested with frequent cutting restriction enzymes which give blunt ends. For example: AluI, HaeIII, and RsaI. Second, the linker was ligated to the blunt ended DNA. Third, the flanking segments were amplified. In this procedure the sample was digested with sufficient enzyme to cut the DNA (about 0.5μL Enzyme) in about 1.5μL of One Phor All Plus Restriction Buffer (Pharmacia) about 250 ng Lambda Phage DNA (50 kb) and sufficient H_2O to bring the final volume to about 15μL. The temperature was held at about 37°C until the DNA was cut.

After the digestion, a cocktail for ligations is added. Although a variety of cocktails are known, the present invention used: up to about 2.0μL of Annealed Oligonucleotides with about 0.05μL of 1M DTT, about 3.0μL of 0.1 M rATP, about 0.25μL of T4 DNA Ligase (Pharmacia), about 1.5μL of One Phor All Plus Restriction Buffer and about 9.7μL of H_2O. The mixture was held at about 15°C for at least about 2 hours. Longer times, for example overnight, may give better results.

Next, the flanking segments were amplified. The amplification mix included about 3μL of STR-PCR Primer

(10 X Cetus; 10μM), about 3μL of Anchor-PCR Primer (same as with STR-PCR primer), about 3μL of dNTP mix (10 X Cetus; 2 mM), about 3μL of PCR Buffer (10 X Cetus, without Mg^{++}), about 3.0μL of 0.01 M $MgCl_2$, about 1μL of DNA, about 14.2μL of H_2O, and about 0.3μL of Amplitaq (Cetus).

The mixture was heated for about 2 min at about 95°C, then the PCR assay on the mixture included about 25-30 cycles of about 45 sec at about 95°C, then about 30 sec at about 60°C, then about 1 min at about 72°C. Finally, the mixture was held at about 72°C for about 10 min, then transferred to about 4°C. In this procedure the STR-PCR was performed separately for both strands of the STR. Control of the concentration of Mg^{++} appeared to be important.

The amplified products were sequenced. If the STR primers are biotinylated (for example, by the Aminolink 2 methodology of Applied Biosystems, Inc.) the products were captured with avidin coated beads (Dynal). The unwanted DNA strand was removed. The preferred conditions for isolation for the amplified products was as follows: about 25μL of M280 Beads from Dynal were mixed with about 25μL of PCR Product. This was held for about 30 min at about 25°C on a rotating wheel. The supernatant was removed and about 150μL of 0.15 M NaOH was added. This was held for 5 min. at about 25°C. The supernatant is then removed, the remaining material is washed at least once with H_2O and resuspended in about 7μL of H_2O for DNA sequencing. Any standard DNA sequencing reactions can be used. In the present invention the sequencing was performed as for any single stranded template.

Example 4

Frequency of Polymorphic Variation of STRs and examples

Seventeen STRs present either within the human HPRT locus or in human sequences in the GenBank database were assayed for variation in the human population. Nine were polymorphic.

Amplifications were performed with Perkin-Elmer-Cetus thermocyclers, Amplitaq enzyme, and recommended buffer conditions in a volume of about 15μL. Amplification conditions were about 95°C for about 45 sec., then about 60°C for about 30 sec., then about 72°C for about 30 sec. Approximately 23-28 cycles were run. Amplified products were radiolabeled by inclusion of 2μCi ^{32}P-dCTP (3000 Ci/mmol) in the PCR. The HUMHPRTB $[AGAT]_n$ and HUMFABP $[AAT]_n$ loci, and the HUMRENA4 $[ACAG]_n$ and HUMTH01 $[AATG]_n$ loci, were studied as a multiplex PCR of two loci. Approximately 50 ng of genomic DNA was used in the PCRs. PCR products were diluted 2:5 in formamide, denatured at about 95°C for about 2 min., and loaded onto a DNA sequencing gel (about 6% (39:1) acrylamide:bisacrylamide, with about 7 M urea, and about 0.04% TEMED). Control reactions without added DNA were included in every set of amplifications. The amplification products ranged in size from between about 100 to 350 bp. This allowed precise determination of allele lengths.

The GenBank data for locus name, approximate repeat sequence, Primer SEQ ID No:, number of alleles observed, number of chromosomes studied and average predicted heterozygote frequencies are shown in Table 6.

TABLE 6: Example STR's

Locus and STR Sequence	PCR Primer SEQ. ID NO:	Alleles Detected	Chromosomes Studied	Heterozygotes (%)	Chromosomes
HUMFABP[AAT]$_{8-15}$	5,6	8	314	52-78	4
HUMARA[AGC]$_{12-30}$	7,8	17	228	87-91	X
HUMGPP3A09[AAGG]$_{9}$	9,10	4	24	31	N.D.
HUMERP[c:AATG]$_{2}$[ACTC]$_{4}$ [c:AATG]$_{5}$	11,12	2	20	19	7
HUMTH01[AATG]$_{6-12}$	13,14	7	320	72-78	11
HUMTNFAB[AATG]$_{5}$	15,16	4	24	31	6
HUMRENA4[ACAG]$_{7-12}$	17,18	6	310	34-48	1
HUMHPRTB[AGAT]$_{9-16}$	19,20	8	227	69-78	X
HUMSTRX1[AGAT]$_{13}$	21,22	11	44	60-84	X

In Table 6 the features of 9 polymorphic STR loci are shown. The range of heterozygote frequencies represents the values obtained for the least to most polymorphic racial group. Alleles from loci shown with a range of reiteration numbers (for example, HUMFABP$[AAT]_{8-15}$) were sequenced to enable precise association of the number of tandem reiterations with specific alleles. The reiteration number of the GenBank clone is given for loci at which the range in the number of repeats is unknown. The lowest alphabetical representation of each STR motif is used, with the reverse complement (c:) indicated where appropriate for compound STR loci. Variability in the human population was assayed with a radioactive PCR assay.

Example 5

Examples of data from the radioactive PCR assay for 5 STR loci

Genotype data for five STR loci were determined in two multiplex and one single PCR (Figure 3). Both DNA strands of the amplified products are radiolabeled and the alleles of different loci have distinct appearances based on the relative mobilities of the two DNA strands. HUMHPRTB $[AGAT]_n$ and HUMTH01 $[AATG]_n$ alleles appear as closely spaced doublet bands, while HUMRENA4 $[ACAG]_n$ and HUMFABP $[AAT]_n$ alleles usually appear as singlets. HUMARA $[AGC]_n$ alleles appear as widely spaced doublets, such that adjacent alleles overlap. The faster strand of HUMHPRTB $[AGAT]_n$ alleles usually appear as a doublet, due to incomplete addition of an

extra, non-complementary base to the 3′ end of the product. The relative mobilities of the strands are influenced by the composition of the polyacrylamide gel. The data for Figure 3 was selected from the population surveys as a fair representation of the clarity with which allele designations were made. The autoradiograms were overexposed to illustrate the faint artifactual bands differing in the number of repeats which are thought to arise during the PCR.

Representative alleles from each of the polymorphic STRs were sequenced. The results show that the variation in size is a function of the number of repeats.

Example 6

Population Genetics of STRs in four human ethnic groups

Trimeric and tetrameric STRs represent a rich source of highly polymorphic markers in the human genome. Analysis of a multilocus genotype survey of 40 or more individuals in U.S. Black, White, Hispanic, and Asian populations at five STR loci located on chromosomes 1, 4, 11, and X was performed. The heterozygote frequencies of the loci ranged from 0.34 to 0.91 and the number of alleles from 6 to 17 for the 20 race and locus combinations. Relative allele frequencies exhibited differences between races and unimodal, bimodal, and complex distributions. Genotype data from the loci were consistent with Hardy-Weinberg equilibrium by three tests and population sub-heterogeneity within each ethnic group was not detected by two additional tests. No mutations were

detected in a total of 860 meiosis for two to five loci studied in various kindreds. An indirect estimate of the mutation rates give values from 2.5×10^{-5} to 15×10^{-5} for the five loci. Higher mutation rates appear to be associated with more tandem repeats of the core motif. The most frequent genotype for all five loci combined appears to have a frequency of 6.51×10^{-4}. Together, these results suggest that trimeric and tetrameric STR loci are ideal markers for understanding the mechanism of production of new mutations at hypervariable DNA regions and are suitable for application to personal identification in the medical and forensic sciences.

TABLE 7: Polymorphic Short Tandem Repeat Loci Studied

Locus and STR	Accession Number	Chromosome	PCR Primer SEQ. ID. NO:	Product Length (bp)	Gene
HUMFABP[AAT]$_n$	M18079	4q31	5,6	199-220	Intestinal Fatty Acid Binding Protein
HUMARA[AGC]$_n$	M21748	Xcen-q13	7,8	261-312	Androgen Receptor
HUMTH01[AATG]$_n$	D00269	11p15.5	13,14	183-207	Tyrosine hydroxylase
HUMRENA4[ACAG]$_n$	M10151	1q32	17,18	251-271	Renin
HUMHPRTB[AGAT]$_n$	M26434	Xq26	19,20	263-299	Hypoxanthine Phosphoribosyl transferase

Samples

DNA was extracted from blood samples obtained at local blood banks from unrelated volunteer donors. Blood bank personnel visually designated donors as Black, White, or Other. Hispanics and Orientals were identified on the basis of surname. A total of 40 individuals in each of these four ethnic groups were studied. Genotype data in 40 families (10 French, 27 Utah/Mormon, 2 Venezuelan, and 1 Amish) was determined with HUMHPRTB $(AGAT)_n$ and HUMFABP $(AAT)_n$. Five STR loci were studied in additional families for a minimum of 31 meiosis.

STR Loci

The STR loci are designated by their GenBank locus name and the lowest alphabetical representation of the 44 possible unique trimeric and tetrameric repeat motifs. For example, HUMHPRTB $(AGAT)_n$ refers to the polymorphic (CTAT) STR located in intron 3 of the human hypoxanthine phosphoribosyltransferase (HPRT) gene. The loci studied, their GenBank accession numbers, chromosomal assignments, amplification primers, and range of product sizes (based on the GenBank sequence) are given in Table 7.

In Table 8 the alleles are numbered according to the number of tandem repeats present in the amplification products. The number of repeated motifs was determined by direct DNA sequencing of amplified products or by subcloning into M13 for sequencing. The repeat number of subcloned fragments was verified

relative to the original genomic DNA source by amplification of the cloned segment.

Computations and Statistics

A variety of standard population genetics tests were employed to evaluate the heterozygote frequencies, allele frequencies and random association of alleles at different loci. These tests included measurements of standard errors, G-statistics for the likelihood-ratio test, binomal distributions, Hardy-Weinberg equilibrium and the summary statistic (S_k^2).

Relative Allele Frequencies

Allele frequencies and their standard errors were calculated from the genotypes of approximately 40 individuals for the 20 combinations of five STR loci and four populations (Table 8).

TABLE 8 Allele Frequencies and Their Standard Errors at Five STR Loci in Four Populations

Allele[a]	Allele frequencies (%) and standard errors (%) in				
	Whites	Blacks	Hispanics	Asians	Pooled[b]
			LOCUS - HUMHPRTB[AGAT]$_n$		
7	0.4±0.4	---[c]	---	---	0.3±0.3
9	0.4± 0.4	1.6±1.6	---	---	0.5±0.4
10	0.4±0.4	1.6±1.6	---	---	0.5±0.4
11	12.1±2.2	3.2±2.2	8.9±3.8	11.3±4.4	10.1±1.5
12	34.4±3.2	29.0±5.8	39.3±6.5	26.4±6.1	33.2±2.4
13	33.0±3.1	30.6±5.9	39.3±6.5	39.6±6.7	34.4±2.4
14	14.7±2.4	21.0±5.2	5.4±3 0	13.2±4.7	14.2±1.8
15	2.2±1.0	11.3±4.0	7.1±3 4	9.4±4.0	5.3±1.1
16	2.2±1.0	1.6±1.6	---	---	1.5±0.6
(n[d])	224	62	56	53	395
			LOCUS - HUMTHO1[AATG]$_n$		
6	26.2±4.9	12.5±3.7	21.3±4.6	8.8±3.2	17.2±2.1
7	8.8±3.2	32.5±5.2	30.0±5.1	23.7±4.8	23.8±2.4
8	11.3±3.5	21.3±4.6	6.3±2.7	3.8±2.1	10.6±1.7
9	16.2±4.1	21.3±4.6	13.7±3.9	47.5±5.6	24.7±2.4
10	36.2±5.4	12.5±3.7	28.7±5.1	7.5±2.9	21.3±2.3
11	1.3±1.2	---	---	7.5±2.9	2.2±2.3
12	---	---	---	1.3±1.2	0.3±0.3
(n)	80	80	80	80	320
			LOCUS -HUMRENA4[ACAG]$_n$		
7	---	2.5±1.7	---	---	0.6±0.5

8	80.3±4.6	71.2±5.1	76.2±4.8	69.7±5.3	74.4±2.5
9	---	3.8±2.1	---	---	1.0±0.6
10	7.9±3.1	11.3±3.5	6.3±2.7	19.7±4.6	11.2±1.8
11	11.8±3.7	10.0±3.4	12.5±3.7	3.9±2.2	9.6±1.7
12	---	1.3±1.2	5.0±2.4	6.6±2.8	3.2±1.0
(n)	76	80	80	76	312
			LOCUS -HUMFABP[AAT]$_n$		
8	---	2.5±1.7	1.3±1.2	---	0.6±0.3
9	0.3±0.3	27.5±5.0	1.3±1.2	5.1±2.5	5.2±1.0
10	49.7±2.9	32.5±5.2	55.0±5.6	66.7±5.3	50.4±2.2
11	17.4±2.2	2.5±1.7	8.8±3.2	6.4±2.8	12.3±1.4
12	3.4±1.0	5.0±2.4	2.5±1.7	2.6±1.8	3.4±0.8
13	24.8±2.5	12.5±3.7	25.0±4.8	19.2±4.4	22.2±1.8
14	4.4±1.2	16.3±4.1	6.2±2.7	---	5.8±1.0
15	---	1.3±1.2	---	---	0.2±0.2
(n)	298	80	80	78	536
			LOCUS -HUMARA[AGC]$_n$		
13	---	1.6±1.6	---	---	0.4±0.4
15	---	---	---	5.7±3.2	1.3±0.8
16	---	1.6±1.6	---	---	0.4±0.4
17	1.7±1.7	17.7±4.9	---	1.9±1.9	5.7±1.5
18	1.7±1.7	16.1±4.7	9.3±3.9	3.8±2.6	7.9±1.8
19	10.2±3.9	8.1±3.5	9.3±3.9	9.4±4.0	9.2±1.9
20	15.3±4.7	9.7±3.8	3.7±2.6	5.7±3.2	8.8±1.9
21	16.9±4.9	8.1±3.5	18.5±5.3	18.9±5.4	15.4±2.4
22	8.5±3.6	17.7±4.9	9.3±3.9	17.0±5.2	13.6±2.3
23	15.3±4.7	4.8±2.7	9.3±3.9	15.1±4.9	11.0±2.1
24	20.3±5.2	6.5±3.1	13.0±4.6	3.8±2.6	10.5±2.0
25	6.8±3.3	1.6±1.6	11.1±4.3	5.7±3.2	6.1±1.6
26	3.4±2.4	---	5.6±3.1	11.3±4.3	4.8±1.4

27	---	1.6±1.6	1.8±1.8	1.9±1.9	1.3±0.8
28	---	3.2±2.2	1.8±1.8	---	1.3±0.8
29	---	1.6±1.6	5.6±3.1	---	1.8±0.9
30	---	---	1.8±1.8	---	0.4±0.4
(n)	59	62	54	53	228

[a]Allelic designations refer to the number of repeats of the core sequence motif indicated in the locus column.

[b]Alleles from all four racial groups were pooled for this column.

[c]A dash indicates the absence of that allele in the respective sample.

[d]Refers to the number of chromosomes samples.

Frequencies of some specific alleles (for example, allele 7 of HUMTH01 $[AATG]_n$ and allele 17 of HUMARA $[AGC]_n$) are clearly variable across the four racial groups. Allele frequency distributions by race are given in Figure 4. With the exception of HUMHPRTB $(AGAT)_n$, which is unimodal and symmetrical, the allele frequency distributions appear bimodal or more complex. The most common allele, however, appears to be the same for some loci (for example, HUMRENA4 $(ACAG)_n$), while at other loci predominant alleles do not coincide between races (for example, HUMARA $(AGC)_n$).

Most Frequent Genotypes

The frequencies of the most common genotypes of a DNA typing system reflect the utility of that assay in practice. The most frequent genotypes for the five STR loci have frequencies from 0.048 to 0.645 in the 20 STR-race combinations (Table 9). The most common genotypes for all five loci combined (p) have

frequencies from 1.40 X 10^{-4} to 6.54 X 10^{-4} in the four racial groups.

TABLE 9 Most Frequent Genotype at the Five STR Loci in Four Populations

Locus	Whites	Blacks	Hispanics	Asians	Pooled
HUMHPRTB[AGAT]$_n$	0.227	0.177	0.309	0.209	0.228
HUMTHO1[AATG]$_n$	0.190	0.138	0.172	0.225	0.118
HUMRENA4[ACAG]$_n$	0.645	0.507	0.581	0.486	0.553
HUMFABP[AAT]$_n$	0.247	0.179	0.303	0.445	0.254
HUMARA[AGC]$_n$	0.069	0.063	0.048	0.064	0.042
Combined (P)	4.74×10^{-4}	1.40×10^{-4}	4.49×10^{-4}	6.51×10^{-4}	1.69×10^{-4}
P^2	2.25×10^{-7}	1.96×10^{-8}	2.02×10^{-7}	4.24×10^{-7}	2.52×10^{-8}

The match probability (P^2) for the most common genotype of all five loci combined was 4.24 X 10^{-7}. The frequencies of the least common genotypes for all five loci combined were on the order of 10^{-17}. The probability calculations in Table 15 are only relevant for female individuals since two of the loci are X-linked. The most common male genotypes for all five loci combined have frequencies from 6.78 X 10^{-4} to 36.4 X 10^{-4} in the four racial groups.

The best markers for individualization in medicine and forensic science may be those with symmetrical and similar allele frequency distributions. Choice of the proper ethnic database appears less critical at such loci for the four ethnic populations we have studied.

The faint artifactual bands which are thought to arise during the PCR, assist in genotype determination relative to external standards. It is possible to count between lanes from allele to allele on overexposed autoradiograms such that even widely separated alleles differing by approximately 6 repeat units were accurately scored. The use of mPCR with fluorescent labels and internal standards improves upon this accuracy.

The data demonstrate that genotype data from trimeric and tetrameric tandem repeats were accurately and efficiently obtained via multiplex PCR. The fidelity with which trimeric and tetrameric STRs were amplified compared to the dimeric STRs make these new class of polymorphic markers well suited for application to DNA typing in forensic science and medicine.

Example 7

Fluorescent DNA profiling assay with internal standards

DNA typing is a powerful technique for determining the relationship, if any, between two genomic DNA samples. Applications for DNA typing include personal identification in paternity testing and forensic science, and sample source determinations in transplantation, prenatal diagnosis, and pedigree validation. Several features of polymorphic STRs suggest that they could form the basis of a powerful and simple DNA typing assay. The small size of the amplified units allows several loci to be easily amplified simultaneously by mPCR, and analyzed with precise allele identification on DNA sequencing gels. The precision, sensitivity, and speed of detecting alleles with PCR offers special opportunities for the study of forensic specimens. For example, trimeric and tetrameric STRs show excellent fidelity of amplification indicating that the genotyping fingerprints can be easily interpreted and are amenable to automation. Fluorescent DNA fragment detection can be used for internal size standards and precise allele quantitation.

For genetic typing, alleles from three chromosomally unlinked STR loci were amplified simultaneously in a mPCR (Figure 5). One primer from each of the three amplification primer sets is differentially labeled with one of the four fluorescent dyes used with the DNA sequencing device. In the ABI 370A system, one dye is reserved for the internal standards, while three dyes are available for the

amplification products of STR loci. Theoretically, any given region of the sequencing gel can contain internal standards as well as alleles from three unlinked STR loci. Used to full potential the approach has enormous personal identification power of high accuracy.

Amplification incorporates a fluorescent label into one end, and a MluI site into the other end of each product in the mPCR (Figure 5). Following amplification of the STR loci from a genomic DNA sample, residual activity of the T. aquaticus polymerase is destroyed and a homogeneous fragment length is achieved for each allele by digestion with MluI. The treated multiplex products are then mixed with internal standards and loaded onto a sequencing gel for analysis on an ABI 370A.

Internal standards were generated by pooling amplification products from individuals of known genotype such that the molar ratios of each allele observed were approximately equal. The pooled alleles were diluted, reamplified, and treated with MluI. This scheme for generating internal standard size markers insures a virtually unlimited supply of standards.

The combination of a quantitative detection system and mPCR enabled additional levels of internal control and precision. Using mPCR products synthesized under standardized amplification conditions, the fluorescent intensity of specific alleles at different loci was related. Because of the relationship between alleles of different loci, it was possible to distinguish between homozygosity and hemizygosity at a given locus (Figure 5). While failure of allele amplification can

occur by primer binding site polymorphism, the null was detectable by quantitation. This quantitative capacity removes the doubt which has been cast on the use of VNTRs due to the observation of homozygosity excess in population studies. The quantitative nature of the allele identification also facilitated the analysis of mixed body samples which occurs in forensics, prenatal diagnosis, the detection of chromosomal aneuploidy, and true somatic mosaicism seen in patients with chromosomal abnormalities and following bone marrow transplantation.

The average individualization potential (P_I) of the three loci together was one in 500 individuals. The combined genotype frequencies (three loci) of the individuals in panels A and B were 0.00026 and 0.0085 assuming Hardy-Weinberg equilibrium. The addition of three more loci will give a P_I of approximately one in 200,000, while the addition of six more loci will give a P_I of one in 90 million. Multiplex PCRs of this complexity have been done. Eight and nine genetic site mPCR for the hypoxanthine phosphoribosyltransferase and dystrophin genes are known.

Oligonucleotides were synthesized on an Applied Biosystems (ABI) 380B DNA synthesizer. Underivatized oligonucleotides were not purified after deprotection and lyophilization. ABI Aminolink 2 chemistry was used to derivatize oligonucleotides for biotin and fluorescent labeling, after which they were ethanol precipitated and purified by polyacrylamide gel electrophoresis. The fluorescent dyes (Molecular Probes, Eugene, OR) used in the assay were (i) NBD aminoheanoic acid for all internal standard markers,

(ii) 5-(and-6)-carboxyfluorescein succinimidyl ester for the HUMTH01 [AATG]$_n$ and HUMHPRTB [AGAT]$_n$ loci, and (iii) Texas Red™ sulfonyl chloride for the HUMFABP [AAT]$_n$ locus. The primer sets had the first primer derivatized and the second primer containing an M1uI restriction site. The primers used were: HUMTH01 (AATG)$_n$, (SEQ. ID NOS: 13, 23); ;HUMFABP(ATT)$_n$ (SEQ ID NOS; 5, 24); HUMHPRTB (AGAT)$_n$ (SEQ ID. NOS; 19, 25). Simultaneous amplification with all six primers was performed with 25 cycles by denaturing at about 95°C for about 45 sec., annealing at about 60°C for about 30 sec., and extending at about 72°C for about 30 sec. using Perkin-Elmer-Cetus thermocyclers, amplitaq, and buffer conditions. The concentration of primers in the multiplex were about 0.06μM for HUMTH01 [AATG]$_n$, about 1.6μM for HUMFABP [AAT]$_n$, and about 0.56μM for HUMHPRTB [AGAT]$_n$. Following amplification, the products were phenol extracted, ethanol precipitated, and digested with M1uI. The digested multiplex products were then combined with the internal size standards and electrophoresed through a 6.5% polyacrylamide, 8.3 M urea gel at 1300-1500 V, 24 mA and 32 W at a temperature of about 46°C. Internal size standards were prepared by amplifying specific alleles from individuals of known genotype. The products were quantitated, combined to give near equimolarity, diluted approximately 5000 fold, and reamplified with approximately 12 cycles.

In Figure 5A, fluorescent profiles of the internal standard cocktails when combined and electrophoresed in a single lane of an ABI 370A DNA sequencing device are shown. In Figure 5B the internal standards were combined with the amplification products of a multiplex

PCR composed of (left to right) the HUMTH01 $(AATG)_n$, (I), HUMFABP $(AAT)_n$ (II), and HUMHPRTB $(AGAT)_n$ (III) loci. The individual shown is heterozygous for all three markets. In Figure 5C multiplex amplification from an individual homozygous at the HUMFABP $(AAT)_n$ locus and hemizygous at HUMHPRTB $(AGAT)_n$ is shown.

Example 8

Alternative detection schemes: radioactive, silver stain, intercalation

Since some forensic laboratories may not have access to fluorescent detection devices, the STR markers can be detected with non-denaturing and denaturing electrophoretic systems using alternative labeling and detection strategies. For example, radioactive and silver staining detection methods, and ethidium bromide staining methods are all applicable. The 6-15 STR loci are sorted into 4 to 5 separate mPCR reactions, each containing 2-4 loci. The three loci are selected such that the amplification products run in non-overlapping regions of the gel (i.e., the base pair lengths of the alleles from different loci do not overlap). Alleles from unknown samples are identified with reference to external standards in adjacent lanes (the same cocktails used in the fluorescent detection scheme (Figure 5) can be employed.

Example 9

Species Specificity

The species specificity of amplification of all STR loci can be determined. Primate DNAs, for example human, baboon, chimpanzee, gorilla, and various bacterial and yeast strains are compared. Additionally, Drosphila, common farm animals, common household pets, and common human flora are also examined. There is no difficulty in obtaining the samples since only 10μg of DNA are needed to perform over 100 studies. The high similarity of sequence between humans and other primates suggests that some of the loci amplify genomes from non-human primates. It is important to document which loci can be amplified from which species for optimal deployment of the method in the forensic arena. Amplification is not seen in non-primates.

Example 10

Kits

The kit includes a container having a oligonucleotide primer pair for amplifying short tandem repeats. The kit can also include standards. One kit includes standards and three oligonucleotide primer pairs. In a preferred embodiment the kit includes sufficient oligonucleotide primer pairs needed to perform mPCR for at least 6-10 loci. The kit can further include the reagents, and established protocols for using the kit. These kits provide for efficient and effective transfer and distribution of the method

to the forensic community. The oligonucleotides and reaction mixtures in these kits can be stored at -70 C for extended periods of time. This facilitates mass production and quality control of the reagents needed to provide accurate reagents at a reasonable cost.

Example 11

Novel STR Sequence

A novel short tandem repeat sequence (SEQ ID. NO. 26) was identified from a lambda clone containing the X chromosome library by screening with a 30 base pair oligonucleotide of the sequence AGAT tandemly repeated. This locus was identified as HUMSTRXI. The sequence flanking the AGAT repeat was amplified and sequenced. Oligonucleotides were designed to amplify the AGAT repeat SEQ ID. NOS. 21 and 22.

The number of AGAT repeats is variable. The exact sequence length was inferred by length polymorphic short tandem repeat sequences with verification of the end sequences. The STR is between approximately base 153 and 203 of SEQ ID. NO. 26. The sense primer is between bases 61 and 84 and corresponds to SEQ ID. NO. 22 and the antisense primer is the reverse complement of the sequence between base 346 and 369 and corresponds to SEQ ID. NO. 21.

Example 12

Results and Benefits Expected

The novel methodology of the present invention provides the most powerful technique to date for the characterization of blood and other body fluids. The increase in credible evidence that this assay produces should result in an increase in the conviction rate for such violent crimes as sexual assault and murder. More importantly, many innocent suspects will be categorically cleared of false accusations. Additional applications would result in increased investigative power in the identification of missing persons, abducted children, military personnel, and human remains from natural and physical disasters. The sensitivity of body fluid identification methods would also be increased well beyond current limits. This would provide obvious benefits in the number of cases in which useful evidence was available.

Another significant improvement over the DNA technology currently used is the profound decrease in the amount of time required to provide results and the amount of labor required to produce the results. The time of actual testing for the new methodology is only 10% of the time required for existing DNA profiling techniques. Furthermore, existing technology is of limited investigative use in sexual assaults and homicides, due to the length of time required to obtain results.

The STR DNA profiling assay enables precise allele determination. The analysis of databases for locus

stability, population heterogeneity, population allele frequencies, the Mendelian inheritance are greatly simplified. The collection of data from the fluorescent STR profiling assay lends itself to automation, thereby reducing the chance of operator error. Defined discrete allele designations promote generation of a national databank.

The loci developed, the profiling assay methods, and the population studies are of interest to the general scientific community. DNA profiling has direct application in the medical diagnostic and research laboratories for verifying specimen identity.

One skilled in the art will readily appreciate that the present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned, as well as those inherent therein. The oligonucleotides, methods, procedures and techniques described herein are presently representative of the preferred embodiments, are intended to be exemplary, and are not intended as limitations on the scope. Changes therein and other uses will occur to those skilled in the art which are encompassed within the spirit of the invention or defined by the scope of the appended claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Albert O. Edwards and Charles Thomas Caskey

(ii) TITLE OF INVENTION: DNA profiling with short tandem repeat polymorphisms and identification of polymorphic STRs

(iii) NUMBER OF SEQUENCES: 26

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Fulbright & Jaworski Patent Department

(B) STREET: 1301 McKinney, Suite 5100

(C) CITY: Houston

(D) STATE: Texas

(E) COUNTRY: U.S.A.

(F) ZIP: 77010-3095

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Disk,3.5 inch (1.44MB)

(B) COMPUTER: IBM PC/AT

(C) OPERATING SYSTEM: MS-DOS

(D) SOFTWARE: BASIC

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER:

(B) FILING DATE:

(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER:

(B) FILING DATE:

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: THOMAS D. PAUL

(B) REGISTRATION NUMBER: 32,714

(C) REFERENCE/DOCKET NUMBER: D-5217

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (713) 651-5325

(B) TELEFAX: (713) 651-5246

(C) TELEX: WESTERN UNION 762829

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 52

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Synthetic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

ACTGCAGAGA CGCTGTCTGT CGAAGGTAAG 40

GAACGGACGA GAGAAGGGAG AG 52

(3) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 52

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Synthetic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

CTCTCCCTTC TCGAATCGTA ACCGTTCGTA 40

CGAGAATCGC TGTCTCTGCA GT 52

(4) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 38

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Synthetic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GCCGGATCCC GAATCGTAAC CGTTCGTACG 30

AGAATCGC 38

(5) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Synthetic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

TACGAGAATC GCTGTCTCTG CAGT 24

(6) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 25

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GTAGTATCAG TTTCATAGGG TCACC 25

(7) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

CAGTTCGTTT CCATTGTCTG TCCG 24

(8) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

TCCAGAATCT GTTCCAGAGC GTGC 24

(9) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

GCTGTGAAGG TTGCTGTTCC TCAT 24

(10) INFORMATION FOR SEQ ID NO: 9:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

TGTGAGTCCC AGTTGCCAGT CTAC 24

(11) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

ACTGGTCACC TTGGAAAGTG GCAT 24

(12) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

TGAGGGCTGT ATGGAATACG TTCA 24

(13) INFORMATION FOR SEQ ID NO: 12:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

CAAGCACCAA GCTGAGCAAA CAGA 24

(14) INFORMATION FOR SEQ ID NO: 13:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

GTGGGCTGAA AAGCTCCCGAT TAT 23

(15) INFORMATION FOR SEQ ID NO: 14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

ATTCAAAGGG TATCTGGGCT CTGG 24

(16) INFORMATION FOR SEQ ID NO: 15:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

GGAGAGACAG GATGTCTGGC ACAT 24

(17) INFORMATION FOR SEQ ID NO: 16:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

CCATCTCTCT CCTTAGCTGT CATA 24

(18) INFORMATION FOR SEQ ID NO: 17:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

AGAGTACCTT CCCTCCTCTA CTCA 24

(19) INFORMATION FOR SEQ ID NO: 18:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18:

CTCTATGGAG CTGGTAGAAC CTGA 24

(20) INFORMATION FOR SEQ ID NO: 19:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 19:

ATGCCACAGA TAATACACAT CCCC 24

(21) INFORMATION FOR SEQ ID NO: 20:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

CTCTCCAGAA TAGTTAGATG TAGG 24

(22) INFORMATION FOR SEQ ID NO: 21:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:

CTCCTTGTGG CCTTCCTTAA ATGG 24

(23) INFORMATION FOR SEQ ID NO: 22:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 24

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:

CTTCTCCAGC ACCCAAGGAA GTCA 24

(24) INFORMATION FOR SEQ ID NO: 23:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 32

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:

TTACGCGTAT TCAAAGGGTA TCTGGGCTCT GG 32

(25) INFORMATION FOR SEQ ID NO: 24:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 32

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:

TTACGCGTCT CGGACAGTAT TCAGTTCGTT TC 32

(26) INFORMATION FOR SEQ ID NO: 25:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 34

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

TTACGCGTTC TCCAGAATAG TTAGATGTAG GTAT 34

(27) INFORMATION FOR SEQ ID NO: 26:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 504

(B) TYPE: Nucleic Acid

(C) STRANDEDNESS: Single

(D) TOPOLOGY: Linear

(ii) MOLECULE TYPE: Genomic DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

TGTTTGTTTT	GTTTTGTTGT	TTTTTTTAAA	TCTGTTCTCA	40
TTGGTGTTTC	TGTTTGCTGC	CTTCTCCAGC	ACCCAAGGAA	80
GTCACCACCA	TATTGTTCCT	TAGTCCTGTG	TTTCTTAGCT	120
GGTCTGCCTT	CTTCTCTCCA	CTTTTCAGAG	TCAGATAGAT	160
GATAGATAGA	TAGATAGATA	GATAGATAGA	TAGATAGATA	200
GATATAAAAA	GATAAATAGA	TAGTCTCTAT	ACAGATATAG	240
ATGTATATCA	TTCCAAGTTT	TAGCTGTATT	TGCAGGAGGA	280
AACGAGAAGT	ATGTCTACTT	CTTTTCCTCT	GGAACCAGAG	320
GTTTCCCCTT	TGTTACCTTG	TTTTTCCATT	TAAGGAAGGC	360
CACAAGGAGA	ACCATGAATC	TGCTCTAATT	GATTTTTACG	400
AAAGGAAGGA	GAAAAAGACA	GGAAATAATT	CACAATCTCC	440
AACTCTTTCA	TCGTAATTAG	TGAGTGACAA	GTAGTTTGTA	480
ACACTCTCAG	TGTATCTTGA	TAAT		504

CLAIMS

WHAT IS CLAIMED IS:

1. A DNA profiling assay for detecting polymorphisms in a short tandem repeat, comprising the steps of:

extracting DNA from a sample to be tested;

amplifying the extracted DNA; and

identifying said amplified extension products for each different sequence, wherein each different sequence is differentially labelled.

2. The method of claim 1, further comprising an external standard.

3. The method of claim 1, further comprising an internal standard.

4. The method of claim 1, wherein the sample to be tested is a forensic or medical sample selected from the group consisting of blood, semen, vaginal swabs, tissue, hair, saliva, urine, and mixtures of body fluids.

5. The assay of claim 1, wherein the short tandem repeat sequence is characterized by the formula $(A_wG_xT_yC_z)_n$ wherein A,G,T and C represent the nucleotides; w,x,y and z represent the number of each nucleotide and range from 0 to 7 and the sum of w+x+y+z

ranges from 3 to 7; and n represents the repeat number and ranges from 5 to 50.

6. The assay of claim 5, wherein the sum of w, x, y and z ranges between 3 to 4 and n ranges between 5 and 40.

7. The assay of claim 1, wherein the short tandem repeat sequence is selected from the non-duplicative alphabetical represented nucleotide sequence group consisting of:

$(AAC)_n$, $(AAG)_n$, $(AAT)_n$, $(ACC)_n$, $(ACG)_n$, $(ACT)_n$ $(AGC)_n$, $(AGG)_n$, $(ATC)_n$, $(CCG)_n$, $(AAAC)_n$, $(AAAG)_n$, $(AAAT)_n$, $(AACC)_n$, $(AACG)_n$, $(AACT)_n$, $(AAGC)_n$, $(AAGG)_n$, $(AAGT)_n$, $(AATC)_n$, $(AATG)_n$, $(AATT)_n$, $(ACAG)_n$, $(ACAT)_n$, $(AGAT)_n$, $(ACCC)_n$, $(ACCG)_n$, $(ACCT)_n$, $(ACGC)_n$, $(ACGG)_n$, $(ACGT)_n$, $(ACTC)_n$, $(ACTG)_n$, $(ACTT)_n$, $(AGCC)_n$, $(AGCG)_n$, $(AGCT)_n$, $(AGGC)_n$, $(AGGG)_n$, $(ATCC)_n$, $(ATCG)_n$, $(ATGC)_n$, $(CCCG)_n$, (CCGG) and combinations thereof; wherein n is the repeat number and varies from about 5 to 20.

8. The assay of claim 1, wherein the DNA is amplified by PCR or multiplex PCR.

9. The assay of claim 1, wherein amplification is by multiplex PCR with primers to at least two short tandem repeat sequences.

10. The assay of claim 1, wherein the label is selected from the group consisting of fluorescers, radioisotopes, chemiluminescers, stains, enzymes and antibodies.

11. The method of claim 10, wherein the label is fluorescent and is selected from the group consisting of Texas Red, NBD aminoheanoic acid Tetramethylrhodamine-5- (and -6) isothiocyanate, and Fluorescein-5-isothiocyanate.

12. The DNA profiling assay of claim 1, further comprising an automated DNA label analyzer capable of distinguishing simultaneously differential labels during the identifying step.

13. A kit for a DNA profiling assay, comprising, a container having oligonucleotide primer pairs for amplifying a short tandem repeat.

14. The kit of claim 9 further comprising a labelled standard.

15. The kit of claim 9 further comprising, a container having reagents for multiplex polymerase chain reaction.

16. A method of detecting a polymorphic short tandem repeat comprising the steps of:

determining possible, non-duplicative nucleotide sequences of the formula $(A_wG_xT_yC_z)$, wherein A,G,T and C represents each nucleotide and w,x,y and z represent the number of each nucleotide and ranges between 0 and 7 with the sum of w+x+y+z ranging between 3 and 7;

searching for $(A_wG_xT_yC_z)_n$ in databases containing known genetic sequences, wherein n

represents the number of tandem repeats of the genetic sequence and is at least 5;

identifying the $(A_wG_xT_yC_z)_n$ sequence and its flanking sequence;

extracting each identified sequence and its flanking sequence;

identifying the extracted sequences which have unique flanking sequences;

17. The method of claim 16, further comprising the steps of:

synthesizing oligonucleotide primer pairs to the unique flanking sequences;

performing a polymerase chain reaction with the primer pairs on DNA samples from a test population; and

examining the extension products of the polymerase chain reaction to detect polymorphic short tandem repeats.

18. A method of detecting polymorphic short tandem repeats comprising the steps of:

synthesizing labelled oligonucleotide probes complementary to the short tandem repeat;

hybridizing the labelled probes to total human λ phage libraries; and

sequencing the hybridized plaques.

19. The method of claim 18, wherein the sequencing step includes subcloning the hybridized plaque.

20. The method of claim 18, wherein the sequencing step includes direct polymerase chain reaction of the hybridized plaque.

21. The short tandem AGAT repeat as defined in SEQ ID. NO. 27.

22. The assay of claim 8, wherein the primer pairs are selected from the group consisting of SEQ ID. NO. 1 and 2, SEQ ID. NO. 3 and 4, SEQ ID. NO. 5 and 6, SEQ ID. NO. 7 and 8, SEQ ID. NO. 9 and 10, SEQ ID. NO. 11 and 12, SEQ ID. NO. 13 and 14, SEQ ID. NO. 15 and 16, SEQ ID. NO. 17 and 18, SEQ ID. NO. 19 and 20, SEQ ID. NO. 21 and 22, SEQ ID. NO. 5 and 24, SEQ ID. NO. 19 and 25 and Seq. ID NO. 13 and 23.

23. The assay of claim 13, wherein the primer pairs are selected from the group consisting of SEQ ID. NO. 1 and 2, SEQ ID. NO. 3 and 4, SEQ ID. NO. 5 and 6, SEQ ID. NO. 7 and 8, SEQ ID. NO. 9 and 10, SEQ ID. NO. 11 and 12, SEQ ID. NO. 13 and 14, SEQ ID. NO. 15 and 16, SEQ ID. NO. 17 and 18, SEQ ID. NO. 19 and 20, SEQ ID. NO. 21 and 22, SEQ ID. NO. 5 and 23, SEQ ID. NO. 19 and 24.

24. The sequences of SEQ ID. NO. 1, SEQ ID. NO. 2, SEQ ID. NO. 3, and SEQ ID. NO. 4 for determing flanking sequences of STR.

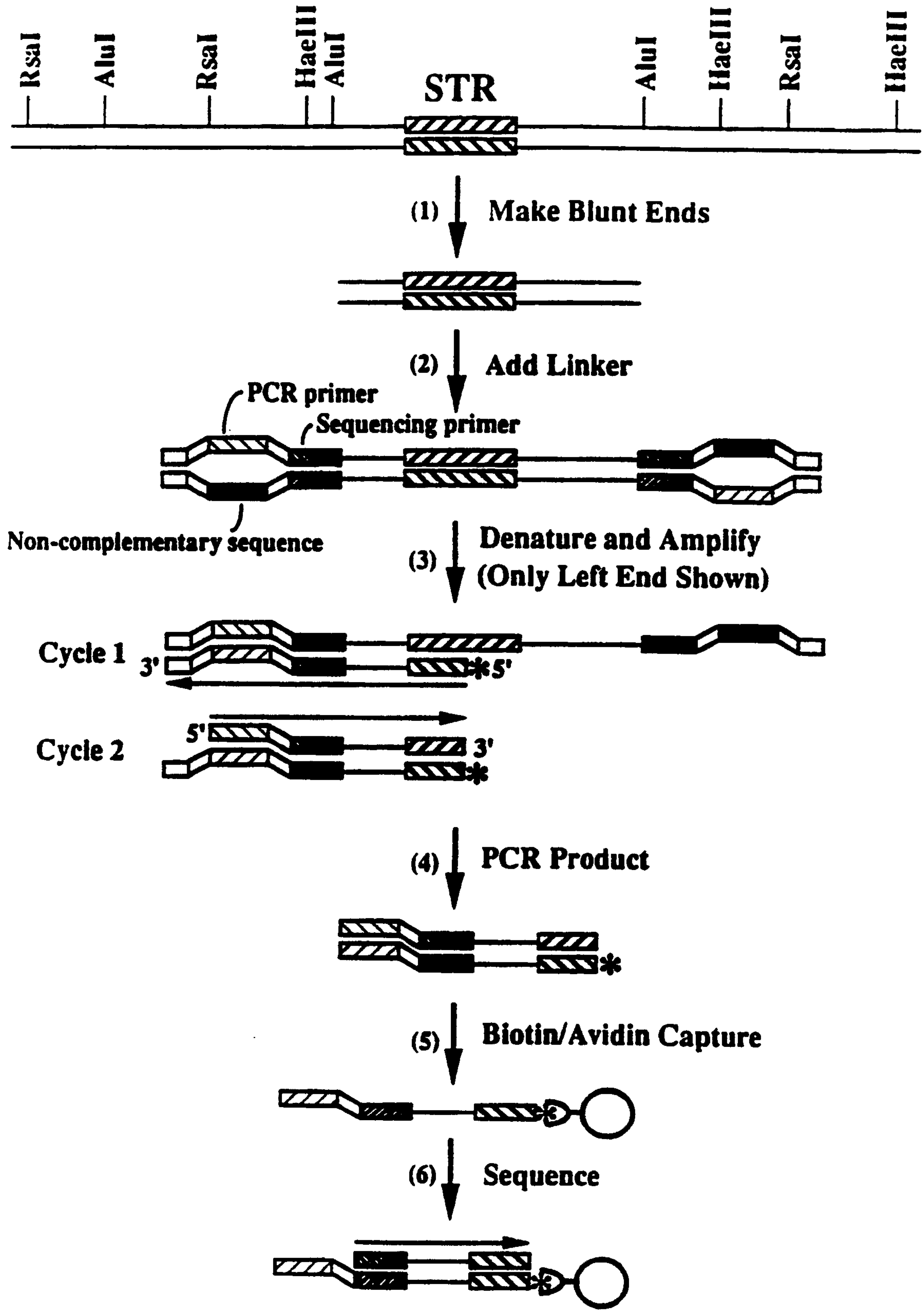

RsaI
AluI
RsaI
HaeIII
AluI
STR
AluI
HaeIII
RsaI
HaeIII
(1) Make Blunt Ends
(2) Add Linker
PCR primer
Sequencing primer
Non-complementary sequence
(3) Denature and Amplify (Only Left End Shown)
Cycle 1
3'
5'
Cycle 2
5'
3'
(4) PCR Product
(5) Biotin/Avidin Capture
(6) Sequence

FIGURE 1

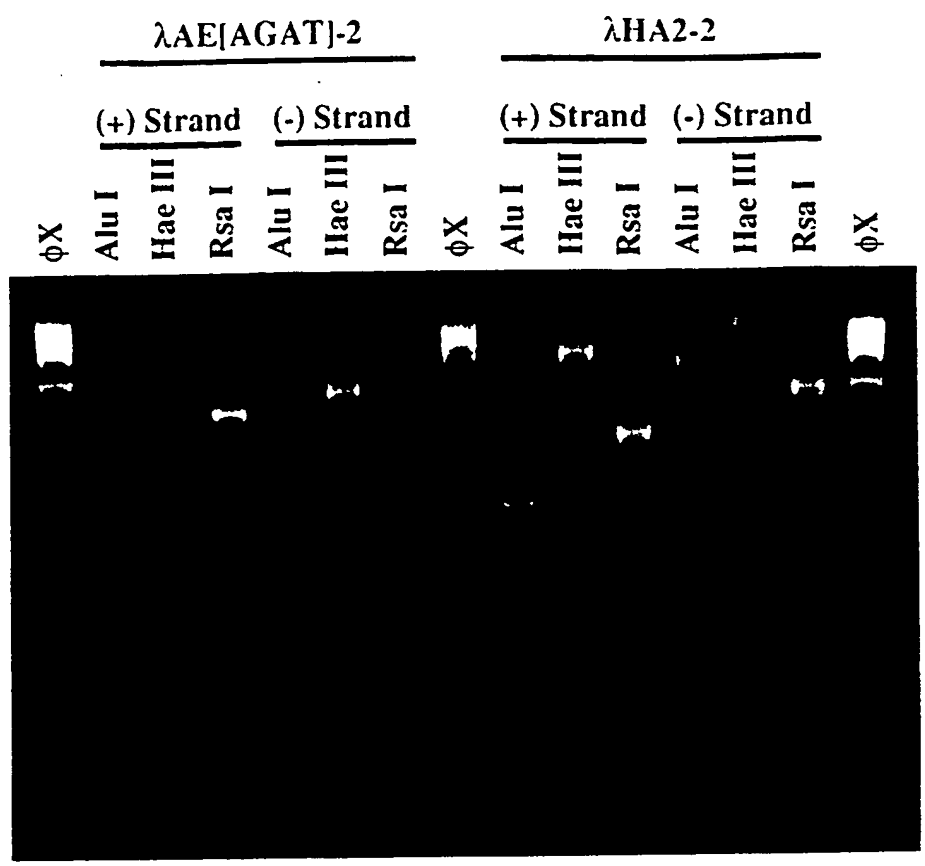


Figure 2A

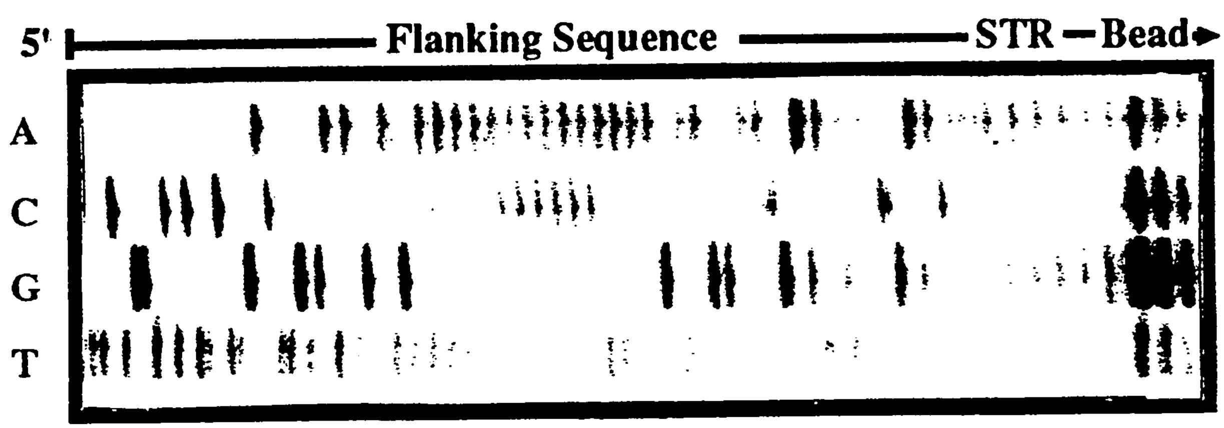


Figure 2B

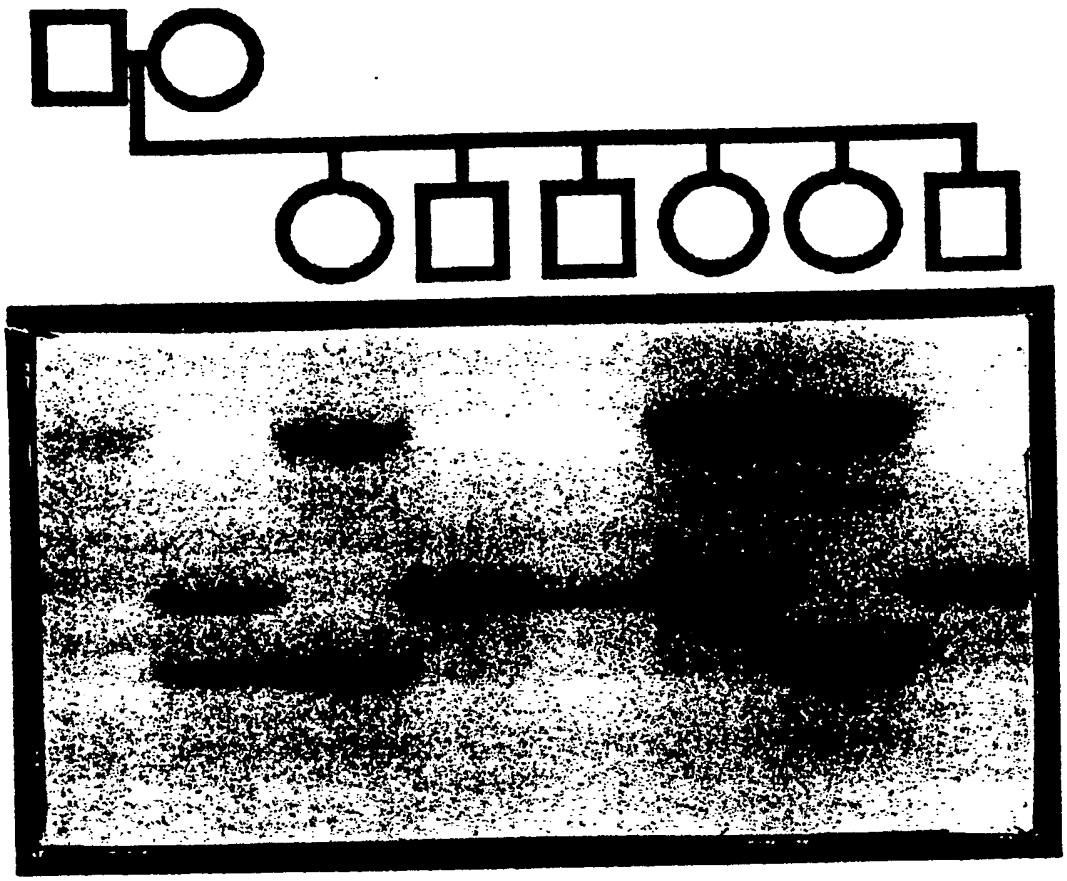


Figure 2C

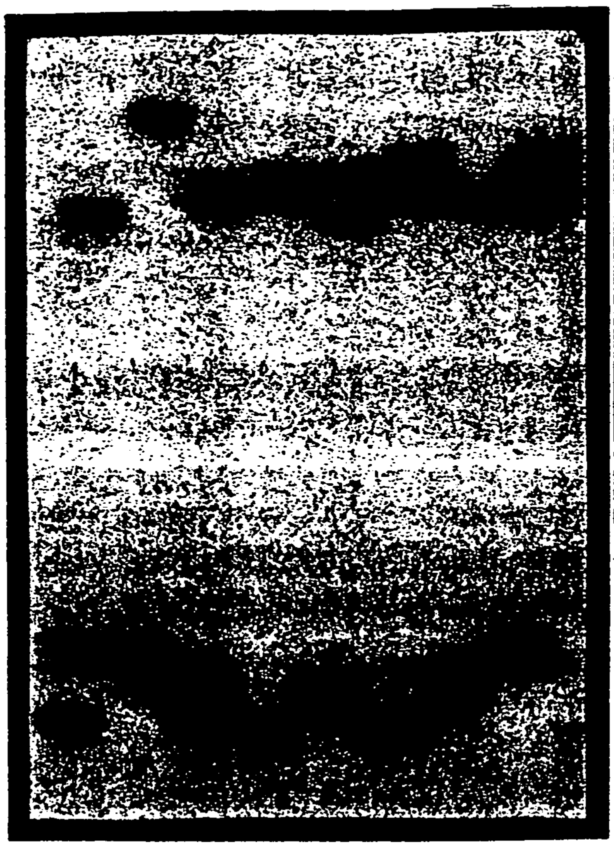


Figure 3A

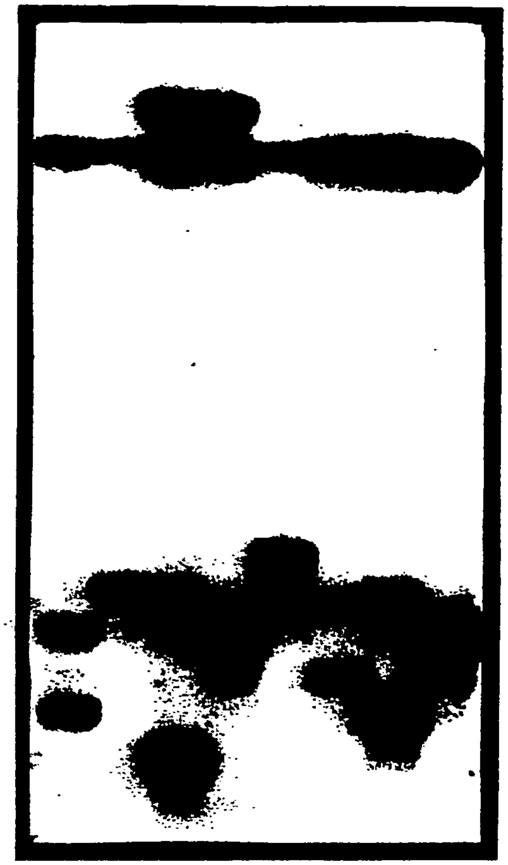


Figure 3B

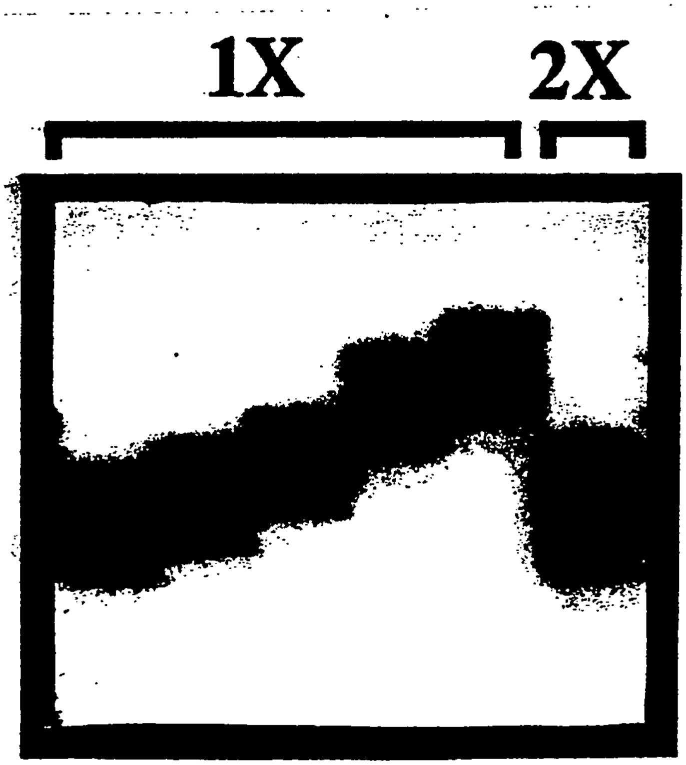


Figure 3C

FIGURE 4A

Frequency Distribution of HUMFABP[AAT] Alleles

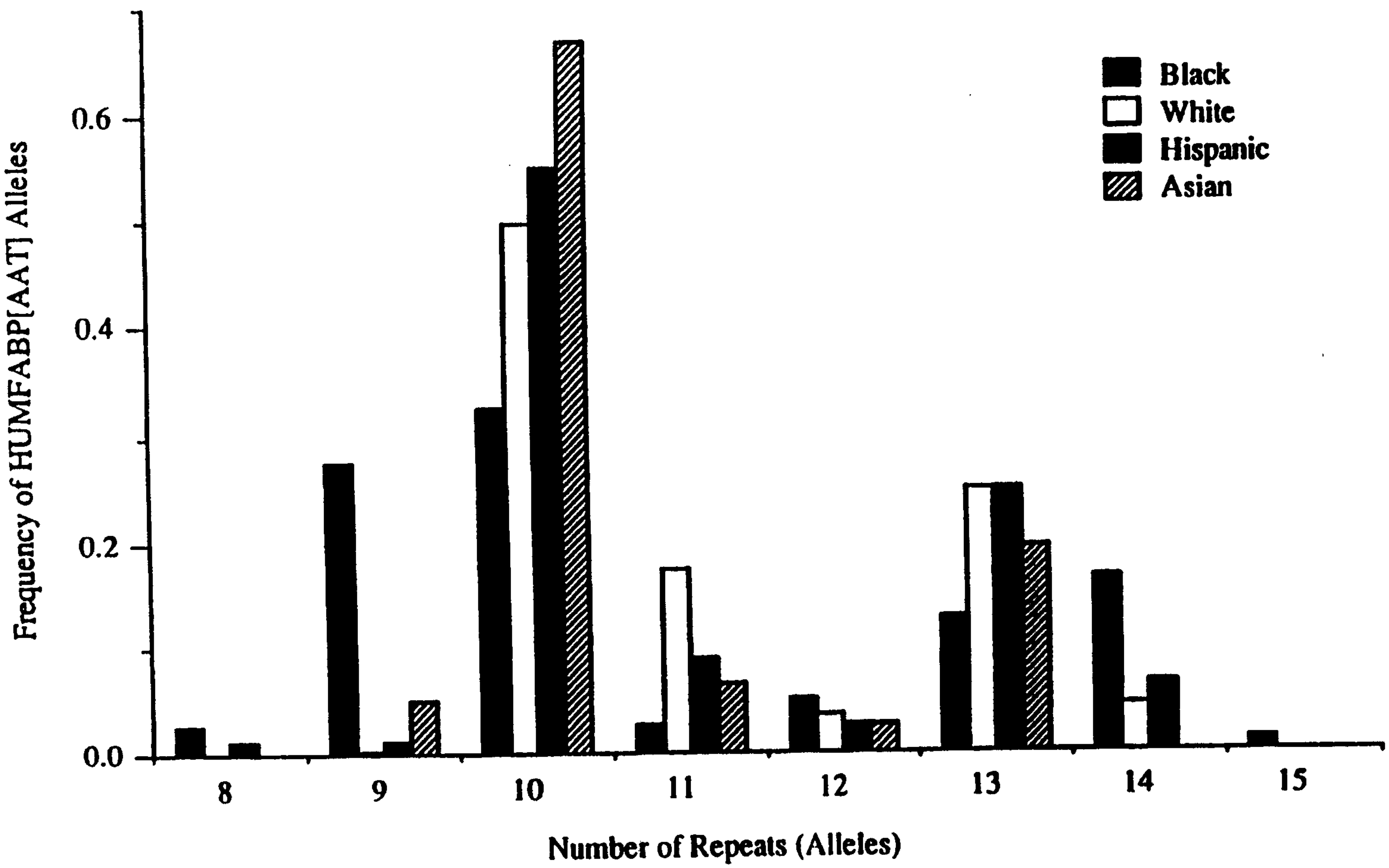


FIGURE 4B

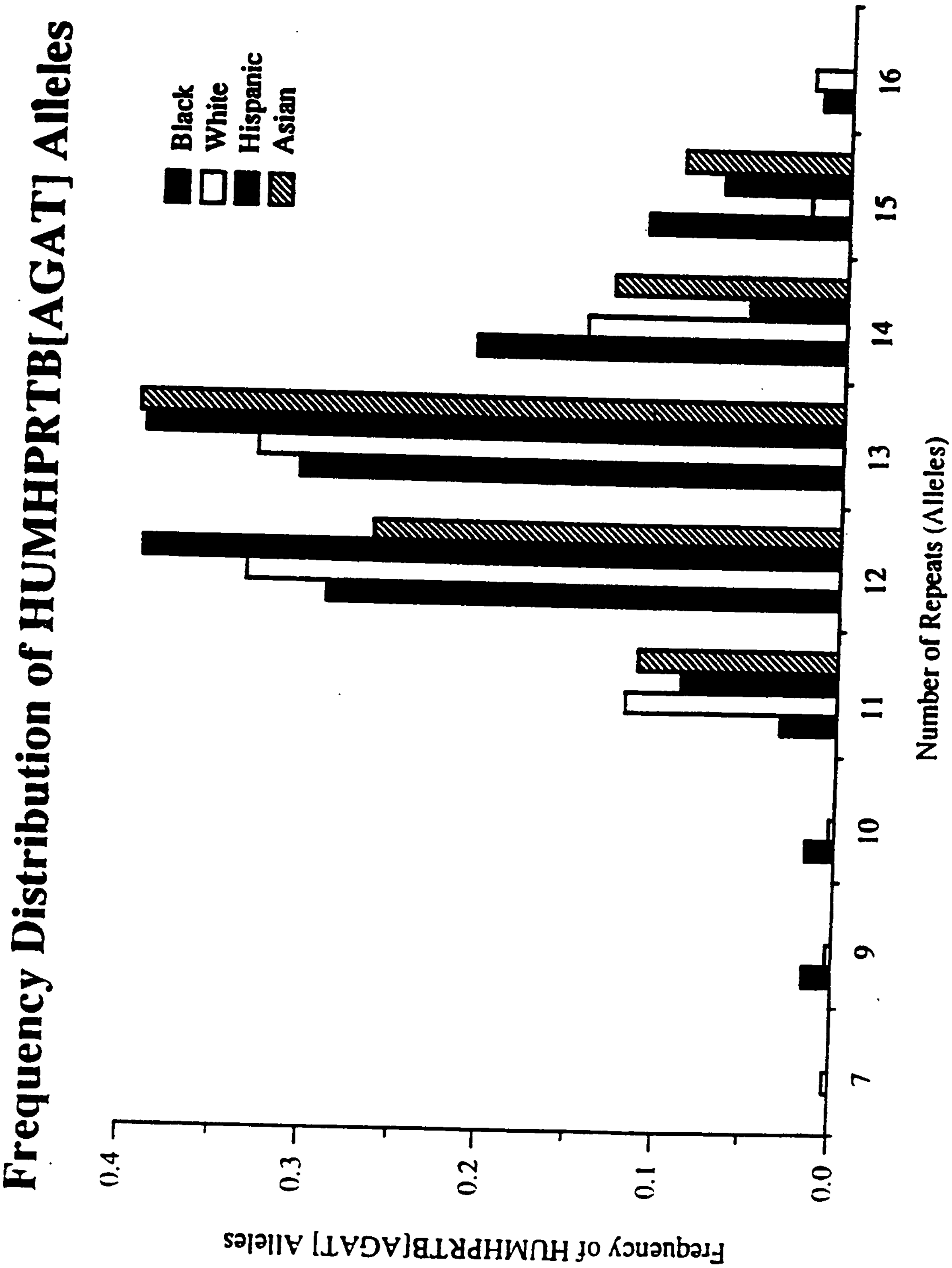

Frequency Distribution of HUMHPRTB[AGAT] Alleles
Black
White
Hispanic
Asian
0.4
0.3
0.2
0.1
0.0
Frequency of HUMHPRTB[AGAT] Alleles
7
9
10
11
12
13
14
15
16
Number of Repeats (Alleles)

FIGURE 4C

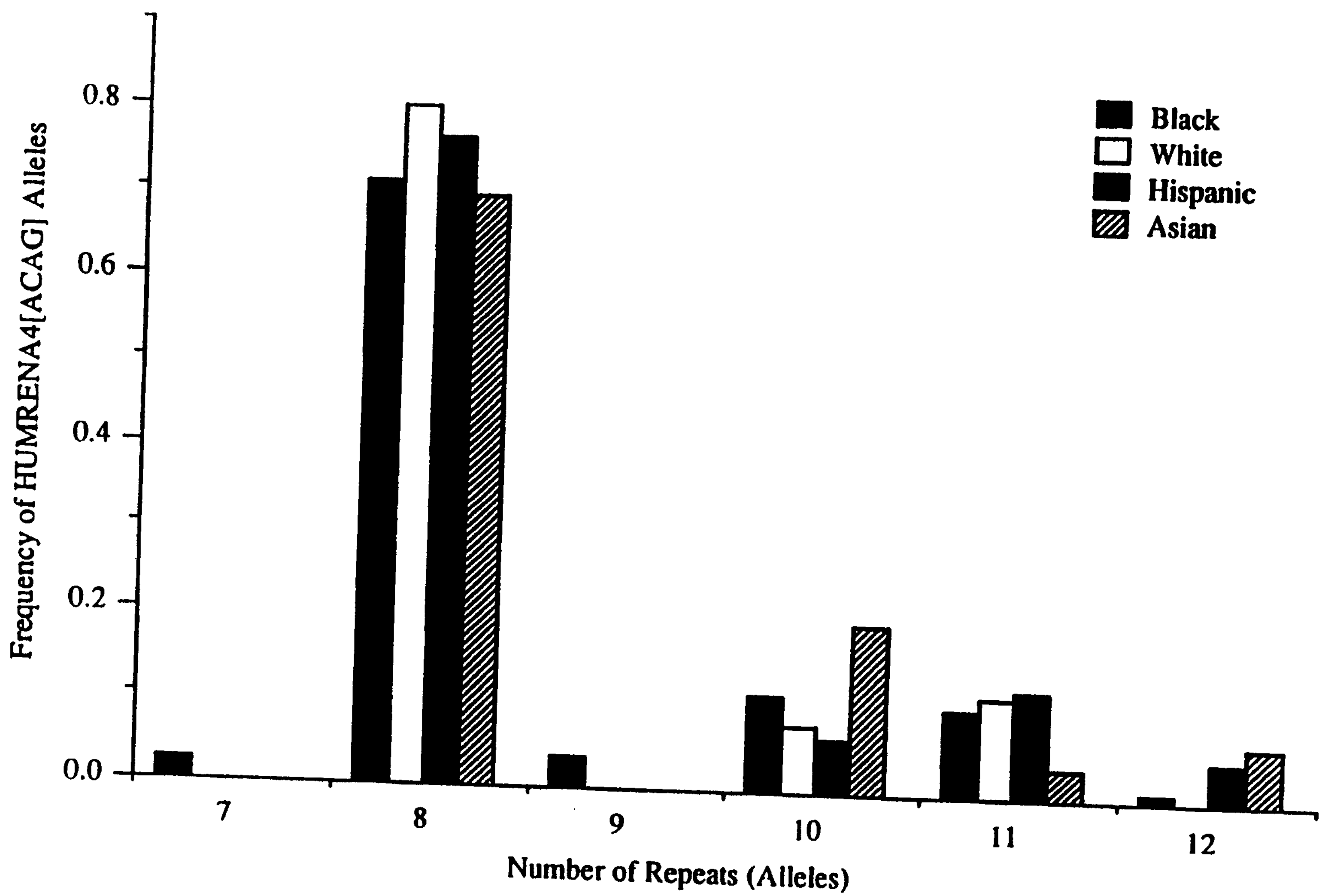
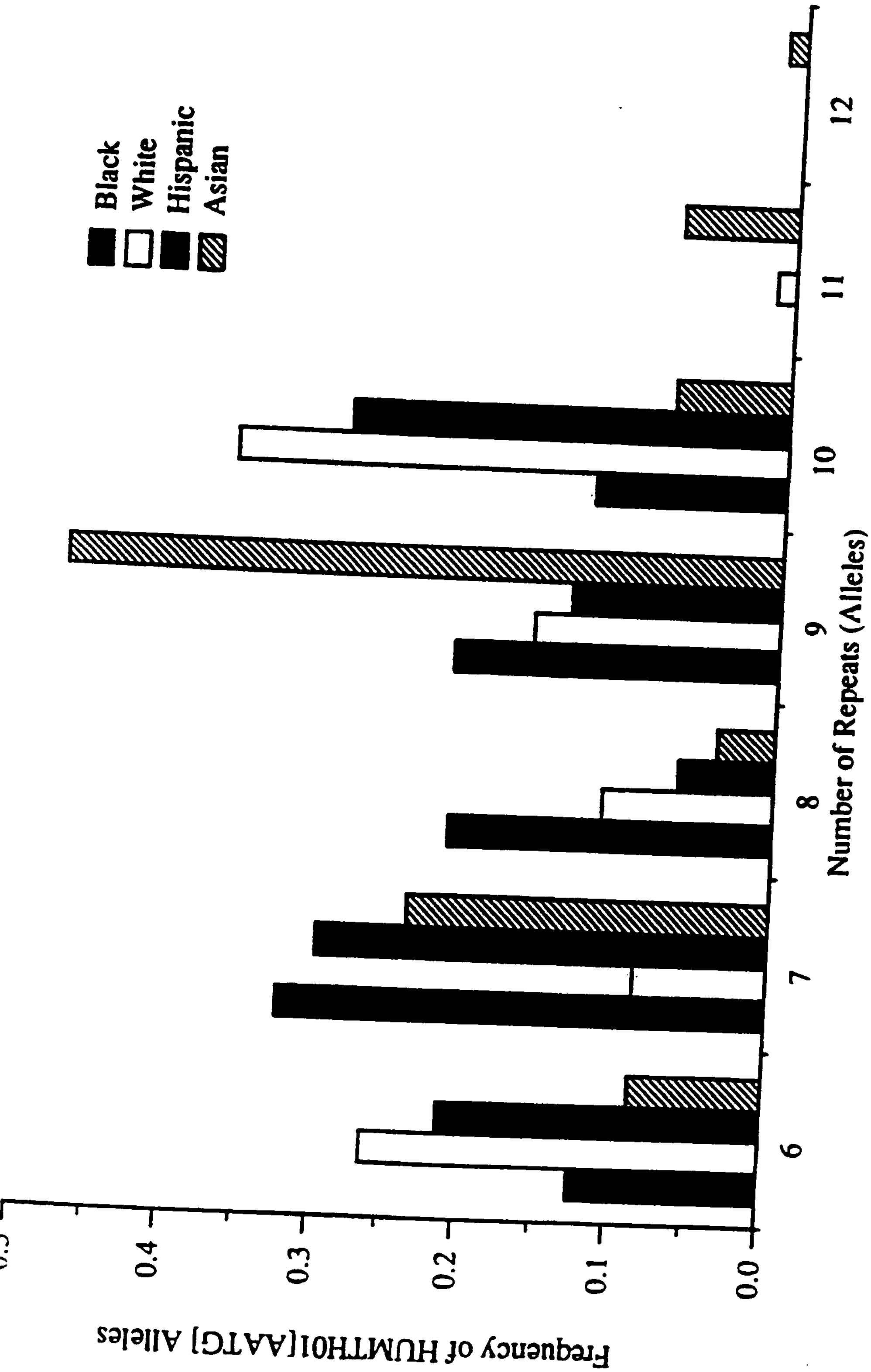

Frequency of HUMRENA4[ACAG] Alleles
0.8
0.6
0.4
0.2
0.0
7
8
9
10
11
12
Number of Repeats (Alleles)
Black
White
Hispanic
Asian

FIGURE 4D


Black
White
Hispanic
Asian
Frequency of HUMTH01[AATG] Alleles
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
Number of Repeats (Alleles)

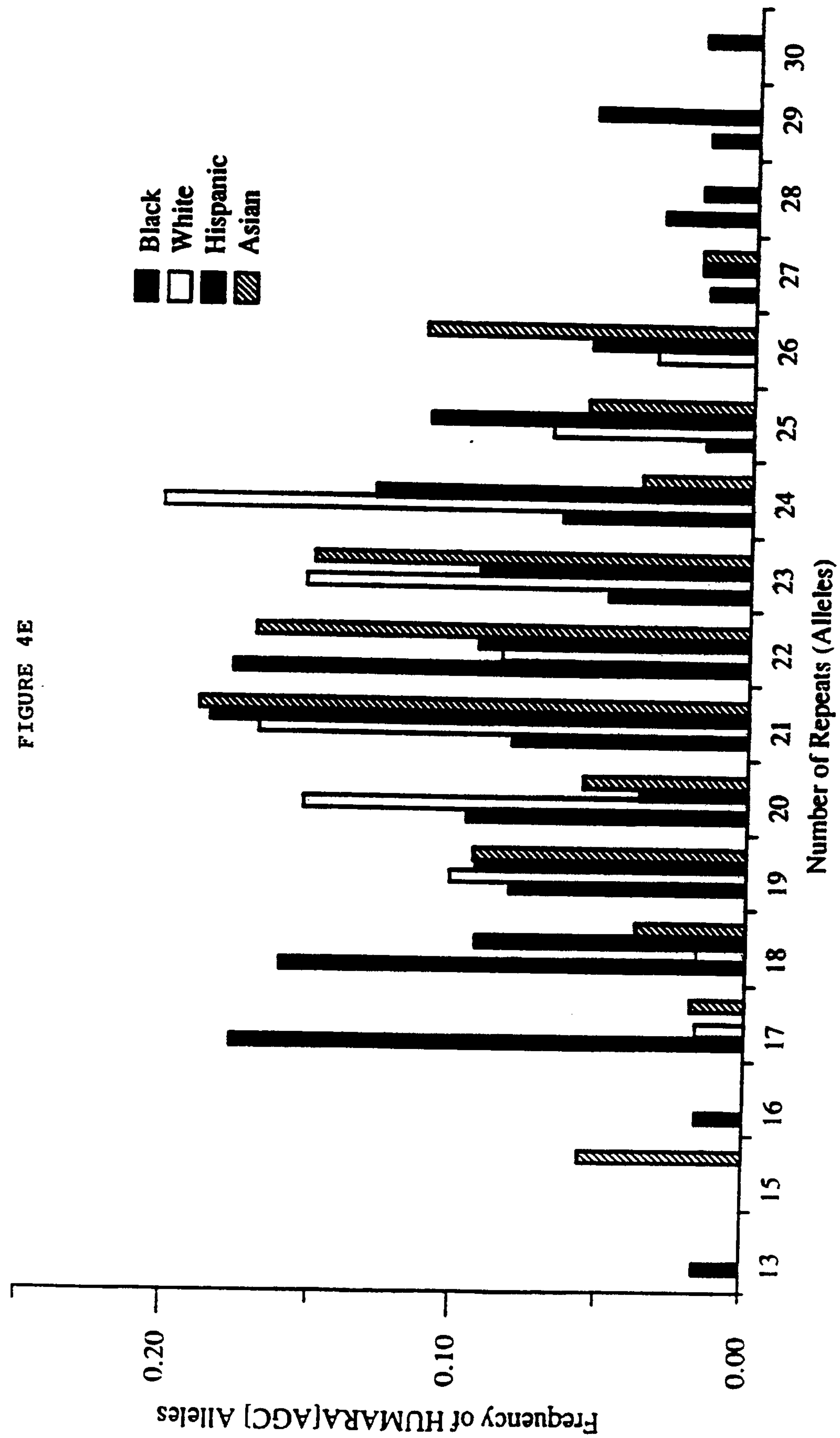

Black
White
Hispanic
Asian
Frequency of HUMARA[AGC] Alleles
0.00
0.10
0.20
Number of Repeats (Alleles)
13
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

FIGURE 4E

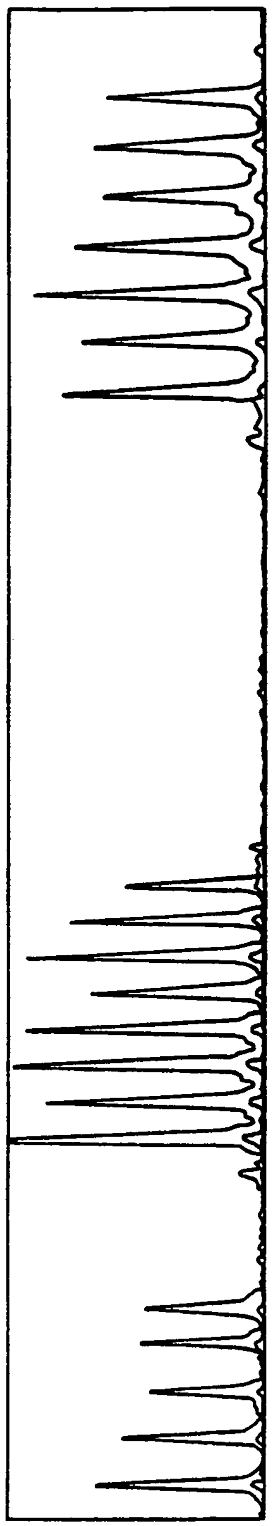


FIG.5A

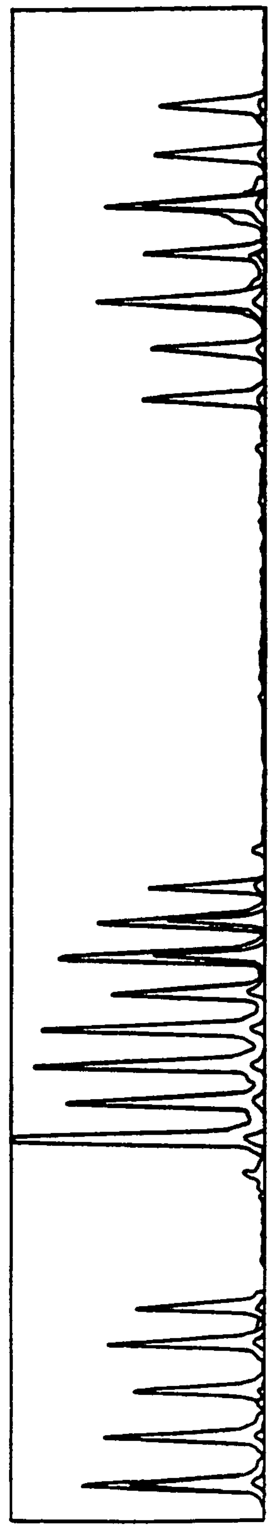


FIG.5B

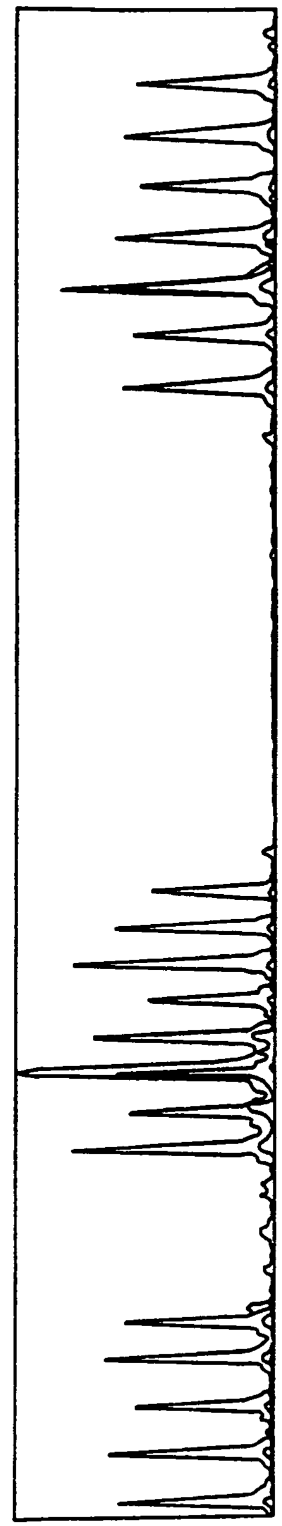


FIG.5C

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all)[3]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC (5): C12Q 1/68; C12P 19/34; C07H 15/12; G01N 33/53
US CL : 435/5, 91; 536/27

II. FIELDS SEARCHED

Minimum Documentation Searched[4]

Classification System	Classification Symbols
U.S.	435/6, 91; 536/27

Documentation Searched other than Minimum Documentation to the extent that such Documents are included in the Fields Searched[5]

APS

III. DOCUMENTS CONSIDERED TO BE RELEVANT[14]

Category*	Citation of Document,[16] with indication, where appropriate, of the relevant passages[17]	Relevant to Claim No.[18]
X/Y	"PCT Technology" published 1989, (New York) pages 209-223, see pages 209-211.	1-4/5-24
Y	Nature, Volume 322, issued 14 August 1986, Tautz et al., "Cryptic simplicity in DNA is a major source of genetic variation, pages 652-656, see abstract.	1-24
Y	Genomics, Volume 7, issued 1990, Gibbs et al., "Multiplex DNA Delection Detection and Exon Sequencing of the Hypoxanthine Phosphorsibiosyltransferase Gene in Lesch-Nythan Families, pages 235-244, see abstract.	1-24

* Special categories of cited documents:[15]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search[2]	Date of Mailing of this International Search Report[2]
15 MAY 1992	05 JUN 1992
International Searching Authority[1]	**Signature of Authorized Officer[20]**
ISA/US	SCOTT CHAMBERS